Image Synthesis in Visual Machine Learning



Fangneng Zhan

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

21/12/2020

.....

Date

Fangneng Zhan

.....

Fangneng Zhan

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

21/12/2020

.....

Date



.....

Prof. Shijian Lu

Authorship Attribution Statement

This thesis contains material from 6 papers accepted at conferences in which I am listed as the first author.

Chapter 3 is related to 3 publications:

Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar Image Synthesis For Accurate Detection and Recognition of Texts in Scenes. European Conference on Computer Vision (ECCV), 2018. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed the experiments, and prepared the manuscript.
- Chuhui Xue and I conducted the experiments.
- The manuscript was revised by Shijian Lu.

Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial Fusion GAN for Image Synthesis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed and conducted the experiments, and prepared the manuscript.
- The manuscript was revised by Hongyuan Zhu and Shijian Lu.

Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Adversarial Image Composition with Auxiliary Illumination. Asian Conference on Computer Vision (ACCV), 2020. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed the experiments, and prepared the manuscript.
- Changgong Zhang and I conducted the experiments.
- The manuscript was revised by Shijian Lu.
- Feiying Ma and Xuansong Xie assisted in the problem formulation.

Chapter 4 is related to 2 publications:

Fangneng Zhan, Shijian Lu. ESIR: End-to-end Scene Text Recognition via Iterative Image Rectification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed and conducted the experiments, and prepared the manuscript.
- The manuscript was revised by Shijian Lu.

Fangneng Zhan, Chuhui Xue, and Shijian Lu. GA-DAN: Geometry-Aware Domain Adaptation Network for Scene Text Detection and Recognition. IEEE International Conference on Computer Vision (ICCV), 2019. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed the experiments, and prepared the manuscript.
- Chuhui Xue and I conducted the experiments.
- The manuscript was revised by Shijian Lu.

Chapter 5 is related to 2 papers:

Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, Xuansong Xie. EMLight: Lighting Estimation via Spherical Distribution Approximation. AAAI Conference on Artificial Intelligence (AAAI), 2021. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed the experiments, and prepared the manuscript.
- Changgong Zhang, Yingchen Yu and I conducted the experiments.
- The manuscript was revised by Shijian Lu and Yuan Chang.
- Feiying Ma and Xuansong Xie assisted in the problem formulation.

Fangneng Zhan, Changgong Zhang, Shijian Lu, Feiying Ma, Xuansong Xie. Sparse Needlets for Lighting Estimation with Spherical Transport Loss. Under Submission. The contributions of the co-authors are as follows:

- I and Shijian Lu proposed the idea, designed the experiments, and prepared the manuscript.
- Changgong Zhang, I conducted the experiments.
- The manuscript was revised by Shijian Lu.
- Feiying Ma and Xuansong Xie assisted in the problem formulation.

21/12/2020 Fangneng Zhan

....

Date Fangneng Zhan

Acknowledgements

I would like to express my sincerest gratitude to my Ph.D. advisor, professor Shijian Lu, who has provided continuous research guidance and enthusiastic encouragement during my Ph.D study. His attitude towards research set a good example of academic researcher for me.

I would also like to thank my collaborators in professor Lu's team. They have provided very insightful advice and feedback on theory and technology for my research.

I wish to express my gratitude towards all the members in MICL Lab for the enthusiastic research atmosphere, stimulating discussions in the last four years.

I also wish to thank the School of Computer Science and Engineering, Nanyang Technological University, for providing the computing resources and technical support.

Last but most importantly, I wish to thank my family and relatives. It is the support from you that drive me to strive for greatness.

Abstract

Image synthesis aims to generate realistic and high-fidelity images automatically. It has attracted increasing interests from both academia and industrial communities in recent years due to its wide applications in various artificial intelligence (AI) tasks as well as the recent advance of generative adversarial networks (GANs). Specifically, image synthesis could generate realistic images of different objects and scenes, hence forms one fundamental component in various design tasks for automated generation of artworks, fashion, advertisement posters, etc. In addition, image synthesis could generate self-annotated images which can be directly applied to deep model training, hence alleviate the data constraint as deep neural networks usually require large amounts of annotated training images that are expensive and time-consuming to collect manually. However, automated generation of realistic and self-annotated images is still facing two major challenges. First, the synthesis realism requires the presence of both image appearance (e.g. image colors, brightness, styles, etc.) and image geometry (e.g. object sizes, alignment, perspective, etc.). Second, generating self-annotated and useful training images is still an open research topic and most existing generation networks cannot handle it well due to the lack of diversity in their generated images.

We investigate automated image synthesis that aims to generate self-annotated and realistic images for either visual design tasks or effective training of deep neural networks. Our works can be broadly grouped into three parts. The first part is composition-based image synthesis that generates realistic and self-annotated images by embedding foreground object into background images automatically. Unlike most existing GAN-based generation networks, it can generate new images with superior diversity as well as new information as the foreground object and background image could come from different sources with completely different distributions. We developed three novel image composition techniques to tackle the challenge of composition-based synthesis that requires to embed foreground objects at the right locations with the right appearance and geometry within the

background image automatically. The second part is translation-based image synthesis that aims to modify existing images to certain new forms that are more useful in deep network training. Unlike most existing image-to-image translation networks that focus on adaptation of image styles and appearance, we developed two image translation networks that adapt image geometries including global image viewpoints and local instance-level object shapes, respectively. The challenge is how to design networks to estimate reliable geometric transformation that often greatly affects the geometry of the translated images. The third part focuses on 3-dimensional (3D) image synthesis, a new problem that aims to embed 3D virtual object models into 2-dimensional (2D) natural images realistically. To ensure that the embedded 3D object models have realistic appearance, we design novel networks that first estimate the environmental lighting of the natural images and then re-light (or render) the embedded 3D objects to have harmonious brightness, color, shadows, etc. with respect to the natural images. Extensive experiments with different evaluation metrics show that our proposed synthesis networks can generate images with superior fidelity and realism. In addition, we demonstrate that our synthesized images are self-annotated and can be directly applied to train deep neural networks for various computer vision tasks effectively.

Contents

Acknowledgements vii

Abstract viii

List of Figures xiv

List of Tables xxii

1 Introduction 1
- 1.1 Scope and Overview . . . 1
- 1.2 Major Contributions . . . 3
  - 1.2.1 Composition-based Image Synthesis . . . 3
  - 1.2.2 Translation-based Image Synthesis . . . 5
  - 1.2.3 3D-Level Image Synthesis . . . 6
- 1.3 Outline of the Thesis . . . 7

2 Literature Review 9
- 2.1 Traditional Image Synthesis . . . 9
  - 2.1.1 Rendering-based Image Synthesis . . . 9
  - 2.1.2 Composition-based Methods . . . 11
- 2.2 Image Synthesis for Data Augmentation . . . 12
- 2.3 Generative Models for Image Synthesis . . . 14
  - 2.3.1 Unsupervised Fundamental Networks . . . 14
  - 2.3.2 Auto-Encoder Networks . . . 15
  - 2.3.3 Generative Adversarial Networks . . . 15

3 Composition-based Image Synthesis 17
- 3.1 Introduction . . . 17
- 3.2 Related Works . . . 19
- 3.3 Verisimilar Image Synthesis for Network Training . . . 22
  - 3.3.1 Semantic Coherence . . . 24
  - 3.3.2 Saliency Guidance . . . 25
  - 3.3.3 Adaptive Text Appearance . . . 27
  - 3.3.4 Experiments . . . 28

3.4 Spatial Fusion GAN for Image Synthesis . . . 34
3.4.1 Geometry Synthesizer . . . 36
3.4.2 Appearance Synthesizer . . . 37
3.4.3 Adversarial Training . . . 40
3.4.4 Experiments . . . 42
3.5 Illumination-Aware Image Composition . . . 47
3.5.1 Local Harmonization . . . 49
3.5.2 Global Harmonization . . . 53
3.5.3 Adversarial Training . . . 53
3.5.4 Experiments . . . 55
3.6 Conclusion . . . 60

4 Translation-based Image Synthesis 63
4.1 Introduction . . . 63
4.2 Related Works . . . 65
4.2.1 Distorted Scene Texts Recognition . . . 65
4.2.2 Domain Adaptation . . . 65
4.3 Iterative Geometry Transformation . . . 66
4.3.1 Rectification Network . . . 68
4.3.2 Recognition Network . . . 71
4.3.3 Network Training . . . 72
4.3.4 Experiments . . . 72
4.4 Geometry-Aware Image Translation . . . 79
4.4.1 Model Architecture . . . 81
4.4.2 Multi-Modal Spatial Learning . . . 82
4.4.3 Adversarial Training . . . 83
4.4.4 Experiments . . . 86
4.5 Conclusion . . . 92

5 3D-Level Image Synthesis 93
5.1 Introduction . . . 93
5.2 Related Works . . . 95
5.2.1 Lighting Estimation . . . 95
5.2.2 Needlets . . . 96
5.3 Earth Mover's Light . . . 96
5.3.1 Regression Network . . . 99
5.3.2 Spherical Mover's Loss . . . 100
5.3.3 Neural Projector . . . 102
5.3.4 Experiments . . . 103
5.4 Sparse Needlets for Lighting Estimation . . . 108
5.4.1 Needlets-based Lighting Representation . . . 110
5.4.2 Sparse Needlets . . . 111
5.4.3 Spherical Transport Loss . . . 113

5.4.4 Experiments 116
5.5 Conclusion 122

6 Conclusion & Future Works **123**
6.1 Conclusion 123
6.2 Future Works 124

List of Author's Publications **126**

List of Figures

1.1 The relationship of different image synthesis approaches. 3

3.1 The proposed scene text image synthesis technique: Given background images and source texts to be embedded into the background images as shown in the left-side box, a semantic map and a saliency map are first determined which are then combined to identify semantically sensible and apt locations for text embedding. The color, brightness, and orientation of the source texts are further determined adaptively according to the color, brightness, and contextual structures around the embedding locations within the background image. Pictures in the right-side box show scene text images that are synthesized by the proposed technique. 23

3.2 Without semantic coherence (SC) as illustrated in (b), texts may be embedded at arbitrary regions such as sky and the head of sheep which are rarely spotted in scenes as illustrated in (c). SC helps to embed texts at semantically sensible regions as illustrated in (d). . . 24

3.3 Without saliency guidance (SG) as illustrated in (b), texts may be embedded across the object boundary as illustrated in (c) which are rarely spotted in scenes. SG thus helps to embed texts at right locations within the semantically sensible regions as illustrated in (d). 26

3.4 Adaptive text appearance (ATA): The color and brightness of source texts are determined adaptively according to the color and brightness of the background image around the embedding locations as illustrated. The orientations of source texts are also adaptively determined according to the orientation of the contextual structures around the embedding locations. The ATA thus helps to produce more verisimilar text appearance as compared random setting of text color, brightness, and orientation. 26

3.5 Several sample images from our synthesis dataset that show how the pro-posed semantic coherence, saliency guidance and adaptive text appearance work together for verisimilar text embedding in scene images automatically. 34

3.6 The proposed SF-GAN is capable of synthesizing realistic images concurrently in geometric and appearance spaces. Rows 1 and 2 show a few synthesized scene text images and row 3 shows a few hat-wearing and glass-wearing images where the foreground texts, glasses and hats as highlighted by red-color boxes are composed with the background scene and face images harmoniously. 35
3.7 The structure of the proposed SF-GAN: The geometry synthesizer is highlighted by blue-color lines and boxes on the left and the appearance synthesizer is highlighted by orange-color lines and boxes on the right. *STN* denotes spatial transformation network, F denotes guided filters, *G1*, *G2*, *D1* and *D2* denote the generators and discriminators. For clarity, cycle loss and identity loss are not included. 36
3.8 Detailed structure of the guided filter F: Given an image to be filtered (*Composed Image* in Fig. 3.7), a translated image with smoothed details (the output of *G1* in Fig. 3.7 where details are lost around background face and foreground hat areas) and the mask of foreground object hat (provided), F produces a new image with full details (*Synthesized Image*, the output of F at the bottom in Fig. 3.7). It can be seen that the guided filter preserves details of both background image (e.g. the face area) and foreground hat (e.g. the image areas highlighted by the red-color box). 39
3.9 Illustration of scene text image synthesis by different GANs: Rows 1-2 are foreground texts and background images as labelled. Rows 3-4 show the images synthesized by ST-GAN and CycleGAN, respectively. Row 5 shows images synthesized by SF-GAN(GS), the output of the geometry synthesizer in SF-GAN (*Composed Image* in Fig. 3.7). The last row shows images synthesized by the proposed SF-GAN. 45
3.10 Illustration of portrait-wearing by different GANs: Columns 1-2 show foreground hats and glasses and background face images, respectively. Columns 3-4 show images synthesized by by ST-GAN [1] and our proposed SF-GAN, respectively. 46
3.11 Given foreground objects and directly composed local images as shows in columns 1 and 2, the proposed AIC-Net disentangles the shadow generation and style transfer to produce realistic local harmonization as shown in column 3. With our spatial transformation module, the locally harmonized images in column 3 can be adapted and composed into the globally harmonized images as shown in column 4. 48

3.12 The structure of the proposed AIC-Net: With a background image *BG* (with specified composition regions) and foreground object *FG* to be composed, a directly composed images *CI* can be derived directly. A shadow branch and a texture branch will learn to generate shadow and transfer the style of *CI* which produces a harmonized *Local Image*. A spatial transformer module (STM) then learns to compose the *Local Image* with the background image to produce a *Global Image*. An *Illumination Model* will predict SH coefficients which can be used to reconstruct an *Illumination Map* ('FC' denote fully connected layer). The *Local Image* will be concatenated with the *Illumination Map* to be distinguished by the *Local Discriminator*. The *Global Image* will be concatenated with an *Embedding Mask* to be distinguished by the *Global Discriminator*. 50

3.13 Illustration of the proposed guided feature filter, branched generation, illumination model and global harmonization: The introduction of the branched generation and guided feature filter helps to suppress various artifacts and generate realistic shadows and foreground object in the local image. The illumination model helps to generate shadow with high-fidelity shadow styles and directions. The global harmonization helps to suppress artifacts and style discrepancy between the local and global images. 51

3.14 Global harmonization: With a selected region in the background image, a homography matrix H can be estimated to transform the selected region to the local image. After the local harmonization, the harmonized local image can be transformed back to the background image by an inverse warping H^{-1}. A masking operation is followed to obtain the global image. 52

3.15 Comparing AIC-Net with other harmonization methods: Images in col. 1 are background images with selected regions. Col. 2 shows directly composed local images. AIC-Net (L) and AIC-Net (G) show images after local and global harmonization. 58

3.16 Shadow generation under different conditions by AIC-Net: The first row shows the background image with the local region for foreground embedding. The second row show the direct composition. Rows 3-4 show the local harmonization and global harmonization by AIC-Net, respectively. AIC-Net can generate various shadows adaptively conditioned on illumination, viewpoint and weather of the background image, etc. 59

4.1 The proposed system consists of an Iterative Rectification Network that corrects scene text distortions iteratively and a Recognition Network that recognizes the finally rectified scene texts (the images within the Iterative Rectification Network illustrate the iterative scene text rectification process). It is end-to-end trainable as driven by better scene text recognition performance, where the dotted lines show the back propagation of gradients. 67

4.2 A novel line-fitting transformation is designed which employs a polynomial to model the middle line of scene texts in horizontal direction, and a set of line segments to estimate the orientation and the boundary of text lines in vertical direction. It helps to model and correct various scene text distortions accurately. 68
4.3 The iterative rectification process: T denotes a thin plate spline transformation, P_1, P_2, ... denote transformation parameters that are predicted by the localization network, Iter denotes the number of rectification iterations and N is the predefined iteration number. 69
4.4 Illustration of scene text rectification and recognition by the proposed ESIR: For the distorted scene text images from the SVTP and CUTE80 in the first column, the second column shows the finally restored scene text images by the proposed rectification network and the third column shows the recognized texts with and without using the proposed rectification technique, respectively. It is clear that the proposed rectification helps to improve the scene text recognition performance greatly. 77
4.5 Visual comparison of different scene text rectification methods: For the four sample images in the first column, columns 2-4 show the rectified images by using the RARE, ASTER, and ESIR, respectively. The sample images are from SVTP and CUTE80 which suffer from perspective and curvature distortions as well as complex image background. The proposed ESIR performs clearly better in scene text distortion rectification. 78
4.6 Domain adaptation by the proposed GA-DAN: For scene text images with clear shifts from the **Source Domain** to the **Target Domain**, GA-DAN models the domain shifts in appearance and geometry spaces simultaneously and generates **Adapted images** with high-fidelity in both appearance and geometry spaces. 80
4.7 The structure of the proposed GA-DAN: S_X (or S_Y) represents the spatial modules as enclosed in blue-color boxes which consist of *Spatial Code*, transformation module T and localization network LN_X (or LN_Y) that predict transformation matrix and transform input images. G_X (or G_Y) denote generators consisting of G_{X_A} (or G_{Y_A}) and G_{X_B} (or G_{Y_B}) as enclosed in green-color boxes that complete the background and translate the image style, respectively. D_X, D_Y and D_T within orange-color boxes denote different discriminators. . 81

4.8 Illustration of the disentangled cycle-consistency loss: S_X, S_Y, G_X and G_Y denotes the spatial modules and generators, respectively, as shown in Fig. 4.7. x, $x_{S_X^{-1}}$ and x_{S_Y} refer to the input images in domain X, reconstructed image by the inverse transformation of S_X and reconstructed image by S_Y. H_{XY} and H_{S_Y} refer to the predicted transformation matrices by S_X and S_Y. ACL and SCL denote appearance cycle-consistency loss and spatial cycle-consistency loss which are obtained by calculating the L1 loss between $(x, x_{S_X^{-1}})$ and (H_{XY}^{-1}, H_{S_Y}), respectively. 84
4.9 Comparing our GA-DAN with state-of-the-art adaptation methods: The first and last columns show source-domain (IC13) and target-domain (IC15) images. GA-DAN_1, GA-DAN_2 and GA-DAN_3 show three GA-DAN adapted images of different spatial views. . . . 90
4.10 Comparing GA-DAN with state-of-the-art adaptation methods: Rows 1-2 show adaptation from COMB to CUTE, Rows 3-4 show adaptation from COMB to SVTP. GA-DAN_1, GA-DAN_2 and GA-DAN_3 show GA-DAN adapted images of different spatial views. 91

5.1 Illustration of our proposed Earth Mover's Light (EMLight): EMLight treats two illumination maps as two discrete spherical distributions. Motivated by the idea of Earth Mover's distances, we design a spherical mover's loss (SML) to measure the distance between two spherical distributions by calculating the minimum distance of moving one distribution to another along the spherical surface. SML aims to find the best *Moving Plan* (with minimum total distance) as illustrated by connections between the two distributions. The thickness of connecting lines denotes the amount of 'Earth' moved between the two points. 97
5.2 Illumination decomposition and estimation in EMLight: The upper and lower graphs illustrate the illumination map decomposition and the structure of the light parameter regression network, respectively. Given an *Illumination Map*, we first derive the *Light sources* region via thresholding and then assign light source pixels to N anchor points as illustrated in *Gaussian Map* (visualized by spherical Gaussian function). We decompose the illumination map into light distribution, light intensity and ambient term and use the decomposition as ground truth for regression network training. The regression network employs a local region (as highlighted by the red box) as the input and three fully connected layers (FC) with output size of $3N$, 3 and 3 (RGB images have 3 channels) to regress the light distribution, light intensity and ambient term, respectively. The estimated illumination parameters are fed to the *Neural Projector* for illumination map generation. 99

5.3 The structure of neural projector: *SConv Block* denote spherical convolution block, F denotes feature fusion blocks, and *Gaussian Map* is acquired through spherical Gaussian mapping according to the predicted parameters. The *Input Image* is fed to an *Encoder* to produce a feature vector for the ensuing spherical generation. The *Gaussian Map* is fused with the multi-scale spherical generation to synthesize the final illumination map. 101
5.4 The scenes used in evaluations consist of three spheres with different materials including diffuse gray, matte silver and mirror silver. . . 104
5.5 Visual comparison of EMLight with state-of-the-art lighting estimation methods: With the illumination maps predicted by different methods (at top-left corner of each rendered image), the rendered virtual objects demonstrate different light intensity, color, shadow and shade. 106
5.6 Illustration of EMLight illumination estimation: For the input images in column 1, columns 2 shows the constructed Gaussian maps based on the regressed illumination parameters and columns 3 and 4 show the generated illumination map under the guidance of Gaussian map and the corresponding intensity map, respectively. Columns 5 and 6 show the ground truth of the illumination maps and the corresponding intensity maps, respectively. 106
5.7 The proposed NeedleLight estimates a parametric lighting representation from a single scene image which is critical to many tasks such as virtual object insertion. Unlike previous methods that predict lighting in either frequency domain [2] (losing spatial localization) or spatial domain [3] (losing frequency information) only, we introduce a novel needlets basis which is capable of representing and estimating lighting accurately in both frequency and spatial domains. 109
5.8 The architecture of the proposed NeedleLight: We introduce sparse needlets to represent lighting. The frequency bands in *Needlet Coefficients* are denoted as $j = 1, \cdots, n$. The needlet coefficients in each frequency band are spatially localized on a set of *Cubature Points* on the unit sphere (illustration for $j = k$ only). For stable lighting regression, we design a spherical transport loss to capture the divergence between the predicted needlet coefficients (*Prediction*) and the ground-truth coefficients (*Ground Truth*). The values of needlet coefficients are illustrated by color bars (only draw 5 for illustration). 110
5.9 The visualization of cubature points (j_{max}=3) on sphere and panorama. The red, green and blue points denote the cubature points of frequency bands from low to high frequency, namely j=1, j=2 and j=3 respectively. 111

5.10 (a) The process to achieve sparse transportation plan through auxiliary points. P and GT are the predicted needlet coefficients and the ground truth, respectively. 'A' denotes the auxiliary point. P′ denotes the predicted coefficients extended with auxiliary points. The connection between P and GT signifies the resulting transportation plan. (b) Transportation plans for the needlet coefficients (j=1) w/o and w/ auxiliary points, respectively. (c) The sparsity and error of transportation plan matrix with different percentage of auxiliary points. 114
5.11 The scene for quantitative evaluations consists of three spheres with diffuse gray, matte silver and mirror silver materials. 116
5.12 Visual comparison of NeedleLight with state-of-the-art lighting estimation methods: With the illumination maps predicted by different methods (at bottom-right corner of each rendered image), the rendered virtual objects demonstrate different lighting intensity, color, shadow and shade. 118
5.13 Object relighting on a variety of photos from the Internet. In all cases, light estimation is performed completely automatically by our model. The predicted illumination maps are utilized to relight the virtual objects with Blender [4]. 119

List of Tables

3.1 Scene text detection recall (R), precision (P) and f-score (F) on the ICDAR2013, ICDAR2015 and MSRA-TD500 datasets, where "*EAST*" denotes the adapted EAST model, "Real" denotes the original training images within the respective datasets, "Synth 1K" and "Synth 10K" denote 1K and 10K synthesized images by our method. 30

3.2 Scene text detection performance on the ICDAR2013 dataset by using the adapted EAST model, where "Synth" and "Gupta" denote images synthesized by our method and Gupta et al. [5] respectively, "1K" and "10K" denote the number of synthetic images used, "Random" means embedding texts at random locations, SC, SG and ATA refer to semantic coherence, saliency guidance, and adaptive text appearance. 31

3.3 Scene text recognition performance over the ICDAR2013, IIIT5K and SVT datasets, where "50" and "1K" in the second row denote the lexicon size and "None" means no lexicon used. "Real" denote the original training images, "Ours 5M", "Jaderberg 5M" and "Gupta 5M" denote the 5 million images synthesized by our method, Jaderberg et al. [6] and Gupta et al. [5] respectively. 33

3.4 Scene text recognition accuracy over the datasets ICDAR2013, ICDAR2015, SVT, IIIT5K, SVTP and CUTE, where 1 million synthesized text images are used for all comparison methods as listed. . . 43

3.5 AMT user study to evaluate the realism of synthesized images. Percentages represent the how often the images in each category were classified as "real" by Turkers. 44

3.6 Comparison of harmonization methods with different metrics: 'Car' and 'Ped' denote the results of car and pedestrian synthesis respectively. 'Local' and 'Global' denote harmonized local and global images. 'MS' metric denotes the manipulation scores. For AMT user study, the number in each cell is the percentage of the harmonized images that are deemed as the most realistic by Turkers. 56

3.7 We conduct ablation study with three metrics (FID, AMT, MS) to evaluate the quality of harmonized local and global images. AIC-Net (WF), AIC-Net (WB), AIC-Net (WS) and AIC-Net(WI) denote AIC-Net without the guided feature filter, without branched generation mechanism, without spatial transformer module and without illumination model respectively. For AMT user study, the percentage numbers represent how often the harmonized images in each category are classified as *most realistic* by Turkers. We average the FID, AMT and MS scores of car and pedestrian images as the final scores. 59

4.1 Scene text recognition performance over the datasets ICDAR2013 (IC13), ICDAR2015 (IC15), IIIT5K, SVT, SVTP and CUTE, where "50" and "1K" in the second row denote the lexicon size and "None" means no lexicon used. The used network backbone and training data are given in [] at the end of each method, where "SK" and "ST" denote the Synth90K and SynthText datasets, respectively. . 75

4.2 ESIR recognition accuracy under different parameter settings: N denotes the rectification iteration number, L denotes the number of line segments, and default setting uses 5 rectification iterations and 20 line segments. When N or L changes, all other parameters keep the same as the default setting (Synth90K and SynthText are used in training under all settings). 79

4.3 Scene text detection over the test images of the target datasets ICDAR2015 and MSRA-TD500: 'IC13', 'Target', 'AD-IC13' and '10-AD-IC13' denote the dataset ICDAR2013, target dataset (ICDAR2015 or MSRA-TD500), 1-to-1 adapted ICDAR2013 and 1-to-10 adapted ICDAR2013, respectively. 'SynthText' refers to 800K synthetic images as reported in [5]. 87

4.4 Scene text detection on the IC15 test images: The detection models are trained using the adapted IC13 training images (from IC13 to IC15) by different adaptation methods as listed. (Baseline is trained by using the original IC13 training images) 89

4.5 Ablation study and comparisons with state-of-the-art adaptation methods: Recognition models are trained by different adaptations of the COMB images to the target domains CUTE and SVTP (Baseline uses the original COMB images and 'Random' applies random spatial transformation in adaptation). 91

5.1 Comparison of EMLight with several state-of-the-art lighting estimation methods: The evaluation metrics include the widely used RMSE, siRMSE, Angular Error (AE) and AMT. D, S, M denote a diffuse, a matte silver and a mirror material of the rendered objects, respectively. 104

5.2 Ablation study of the proposed EMLight: SG and SD denote the spherical Gaussian representation and our spherical distribution representation of the illumination map. L2 and SML denote using the L2 and spherical mover's loss to regress the representation parameters. NP denotes the proposed neural projector. 107
5.3 Ablation studies over anchor points, loss functions and convolution types: EMLight* denotes the standard EMLight with 128 anchor points, spherical mover's loss (SML), and spherical convolution. We create four EMLight variants by setting the number of anchor points to 64 and 196, replacing SML with cross-entropy loss, and replacing spherical convolution with normal convolution. 108
5.4 Comparison of NeedleLight with several state-of-the-art lighting estimation methods: The evaluation metrics include the widely used RMSE, si-RMSE, AMT, and our proposed STD. D, S, M denote a diffuse, a matte silver and a mirror material of the rendered objects, respectively. 117
5.5 Ablation study of the proposed NeedleLight: *SH* denotes using spherical harmonics for representation; *SN* denotes using spherical needlets for representation; *ST* denotes applying the derived soft thresholding function to needlet coefficients; *L2* and *STL* denotes regressing coefficients with L2 loss and spherical transport loss; *SN+ST+L2+STL* denotes standard NeedleLight with sparse needlets which are regressed with both L2 and spherical transport loss. 120
5.6 Ablation studies over different bases including spherical harmonics (SH), spherical Gaussian (SG), Haar, spherical needlets (SN). HT and ST denote apply hard thresholding and the derived soft thresholding functions to the spherical needlet coefficients. j_{max} denotes the order of needlets for representation. SN(j_{max}=3)+ST* is the standard setting of NeedleLight. 121

Chapter 1

Introduction

1.1 Scope and Overview

Visual data plays a significant role in our daily activities as we consume astounding amounts of visual contents such as photos and videos every day. Learning to generate realistic visual data will help us to understand and develop machine learning models for perceiving the visual world. On the other hand, automated generation of realistic visual data remains a grand challenge due to the super-rich complexity and variations of visual data. Realistic synthesis of visual data such as images has been a hot research topic for a long time in the computer vision and computer graphic community thanks to its wide applications in advertisements, gaming, virtual reality, etc. With the increasing demands of deep learning applications, image synthesis also demonstrates very good potential for becoming a vital component for generating useful training data as required in the deep network training pipeline. This section provides an overview of the developments of image synthesis methods, starting from the traditional image synthesis, image synthesis for machine learning and the recent adversarial image synthesis. Specifically, we start with an overview of traditional image synthesis and its introduction in the machine learning community, followed by learning-based synthesis methods especially generative networks for image synthesis.

Automatic generation of image contents has been a research topic since the mid of 20th century. With the works on fundamental concepts such as ray tracing [7] and shading [8, 9], this topic enter the stage of rapid development. After that,

the development of this area was further driven by the interest in film industry, advertisements and computer games. In the 1980s, the optical flow research in computer vision community demands annotated data for evaluations [10–12], while it is almost impossible to annotate the flow vectors by manually. Thus, synthetic data starts to play significant roles in optical flow and other computer vision tasks.

With the emergence and fast development of deep learning in recent years, many research areas including computer vision have considered deep learning as an integral part to reach superior performance. As a data-driven learning approach, the requirement of large amounts of annotated training data has become one bottleneck for wide applications of deep learning. To relieve this data constraint, synthetic data has attracted increasing interest in recent years as a data augmentation technique in deep network training. For example, Karsch et al. [13] develop a semi-automatic technique to insert objects into photographs with realistic perspective and illumination; Aldrian et al. [14] propose an inverse rendering approach for synthesizing a 3D structure of the face; Dosovitskiy et al. [15] apply synthetic floating chair images to train optical flow networks; Jaderberg et al. [6] create a synthetic text-image dataset to train the text recognition models; Yildirim et al. [16] utilize the deep features trained on synthetic faces to regress face pose parameters.

Synthetic data can greatly drive the advance of deep learning for training deep networks effectively and efficiently. Inversely, the advances in deep learning also pushes the progress of realistic image synthesis. Specifically, Generative Adversarial Networks [17] have achieved great success in generating realistic images from either random noises or some prior guidance. The main idea is to conduct an adversarial learning between a generator and a discriminator, where the generator aims to generate more realistic images to fool the discriminator which aims to distinguish the generated images from real images. The performance of generative models have been improved greatly by the Laplacian pyramid of adversarial networks [18] and the following equipment of a DNN architecture [19], learning more interpretable latent representations [20], adopting an alternative training [21]. In addition, GAN-based image-to-image translation has been investigated extensively due to its wide applications in different tasks such as domain adaptation [22], high resolution image synthesis [23], image inpainting [24], etc. Existing works explored

different conditional inputs such as semantic segmentation [25–27], scene layouts [28–30], key points [31, 32], edge maps [25], etc.

1.2 Major Contributions

The major contributions of my PhD research lie with the exploration of images synthesis from different perspectives. My research can be grouped into three categories including composition-based image synthesis, translation-based image synthesis and 3D-level image synthesis as shown in Fig. 1.1. More details to be described in the ensuing subsections.

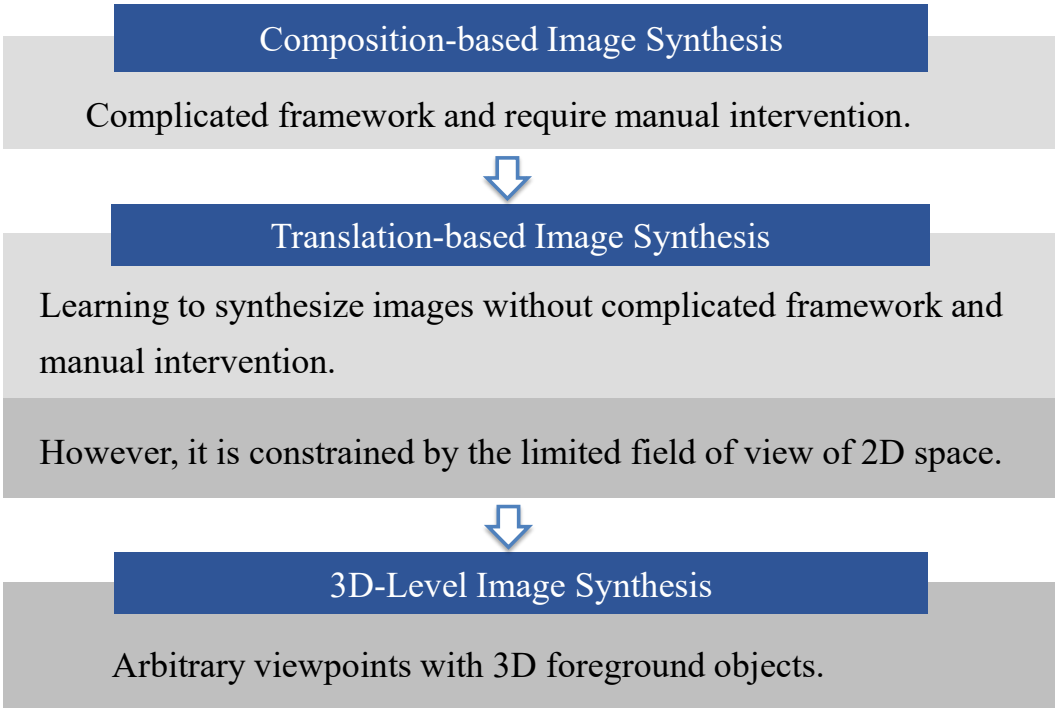


FIGURE 1.1: The relationship of different image synthesis approaches.

1.2.1 Composition-based Image Synthesis

The objective of this research is to explore image synthesis by composing foregrounds objects with background images. We explore three innovative approaches for achieving efficient and effective image composition.

We first studied a traditional image composition approach that aims to generate a large amount of realistic scene text images for training accurate and robust scene text detection and recognition models. To make the synthesized scene text images realistic and useful in network training, the composed images should be as close to what they really look like in natural scenes as possible. We therefore designed a semantic coherent module and a visual saliency module for determining the right embedding locations of the foreground objects within the background images. Additionally, we introduced an adaptive appearance module to determine realistic appearance for the embedded foreground texts. To ensure that the generated scene text images have sufficient diversity, our image synthesis approach controls the variation of the generated images from two aspects: 1) it straightly applies different geometric transformations and alternations of image colors and brightness to introduce variations to the foreground texts; 2) it employs a large number of natural images of different characteristics to introduce variations to the background images. Extensive experiments demonstrate the effectiveness of this proposed image synthesis approach - better scene text detection and recognition networks were trained while incorporating our synthesized images.

Although our traditional image composition method can synthesize training image effectively, the realism of the synthesized images still has clear gaps as compared with natural images. As generative adversarial networks (GANs) have shown great potentials in realistic image synthesis, we designed an innovative Spatial Fusion GAN (SF-GAN) [33] that employs a geometry synthesizer and an appearance synthesizer to synthesize images with realism in both geometry and appearance spaces. The geometry synthesizer learns contextual geometries of background images to spatially transforms the foreground objects to be matched with the background images. The appearance synthesizer adjusts the color, brightness and styles of the foreground objects to ensure the harmonization with the background images, where a guided filter is employed for detail preservation. The geometry synthesizer and appearance synthesizer can be trained end-to-end without supervision and serve as mutual references in training. Extensive experiments over several tasks demonstrate the superior performance of SF-GAN for image synthesis as compared with the state-of-the-art.

State-of-the-art image composition methods strive to harmonize the composed images by adapting the style of foreground objects to be compatible with that of the

background images, whereas the potential shadow of foreground objects within the composed images which is critical to the composition realism is largely neglected. We proposed an Adversarial Image Composition Net (AIC-Net) [34] that achieves realistic image composition by considering potential shadows that the foreground object projects in the composed images. A novel branched generation mechanism was designed, which disentangles the generation of shadows and the transfer of foreground object styles for optimal accomplishment of the two tasks simultaneously. A differentiable spatial transformation module is designed which bridges the local harmonization and the global harmonization to achieve their joint optimization effectively. Extensive experiments over pedestrian and car image composition tasks show that the proposed AIC-Net achieves superior composition performance qualitatively and quantitatively.

1.2.2 Translation-based Image Synthesis

Image-to-image translation aims to adapt image styles and appearance and it has been widely adopted to mitigate domain gaps in training domain adaptive models. On the other hand, most existing translation networks focus on the adaptation of image styles and appearance where the inter-domain discrepancy in image geometry is largely neglected. We explored how to translate images in geometry space by adapting either global image viewpoints (up to the camera view angle) or local object shapes and perspectives.

Objects in scenes often suffer from perspective distortions that tend to degrade object recognition greatly. We designed a geometry translation framework ESIR [35] for the rectification of perspective distortions of texts in scenes. We trained ESIR in an unsupervised manner as driven by the performance of a scene text recognition module. The network can iteratively transform perspective-distorted texts with a novel line-fitting transformation which estimates the pose of text lines in scenes and corrects it towards horizontal. ESIR is robust to parameter initialization and the training needs scene text images and word-level annotations only. Extensive experiments over a number of public datasets shows ESIR achieves superior instance-level image translation in the geometry space by transforming scene texts from perspective views to straight and horizontal views.

ESIR focuses on the translation of instance-level geometries only, where realistic image translation requires to translate both image geometry and image appearance simultaneously. We thus explored to achieve image translation in both geometry and appearance domains under the context of domain adaptation. Specifically, we designed an innovative Geometry-Aware Domain Adaptation Network (GA-DAN) [36] that can model cross-domain shifts in both geometry space and appearance space concurrently and convert images across domains with very different characteristics realistically. In GA-DAN, we designed a novel multi-modal spatial learning technique that converts a source-domain image into multiple images of different spatial views with similar target-domain appearance and styles. We introduced a new disentangled cycle-consistency loss that balances the cycle consistency in appearance and geometry spaces and improves the learning convergence and stability of the whole network effectively. The proposed GA-DAN has been evaluated for the classic scene text detection and recognition tasks, and experiments show that the domain-adapted images achieve superior scene text detection and recognition performance while applied to network training.

1.2.3 3D-Level Image Synthesis

Image synthesis in 2-dimensional (2D) space often faces various view constraints. Take image composition as an example. Image composition with 2D foreground images can only generate new images with a fixed viewpoint for the embedded foreground objects. Image synthesis in 3-dimensional (3D) space can greatly mitigate such constraints by embedding 3D foreground object models in various poses and views. We investigated image synthesis with 3D object models. In particular, our studies focus on how to estimate the lighting of the background image so as to generate harmonious appearance (in color, brightness, shadows, etc.) for the embedded 3D foreground object models.

To embed 3D objects into 2D background images realistically, we need to recover the illumination of the background image so as to re-light/render the 3D objects harmoniously. Recovering illumination from a single image has been investigated extensively in the computer vision and computer graphic research community. On the other hand, existing works estimate illumination by either regressing light parameters or generating illumination maps that are often hard to optimize or tend

to produce inaccurate predictions. We designed Earth Mover's Light (EMLight) [37], an illumination estimation framework that leverages a regression network and a neural projector for accurate illumination estimation. We decompose the illumination map into spherical light distribution, light intensity and the ambient term, and define the illumination estimation as a parameter regression task for the three illumination components. Motivated by the Earth Mover's distance, we designed a novel spherical mover's loss that guides to regress light distribution parameters accurately by taking advantage of the subtleties of spherical distribution. Under the guidance of the predicted spherical distribution, light intensity and ambient term, the neural projector synthesizes panoramic illumination maps with realistic light frequency. Extensive experiments show that EMLight achieves accurate illumination estimation and the generated relighting in 3D object embedding exhibits superior plausibility and fidelity as compared with state-of-the-art methods.

In addition, existing approaches model lighting in either frequency domain or spatial domain which is insufficient to represent the complex lighting conditions in scenes and tends to produce inaccurate estimation. We therefore designed NeedleLight [38], a new lighting estimation model that represents illumination with needlets and allows lighting estimation in both frequency domain and spatial domain jointly. An optimal thresholding function is designed to achieve sparse needlets which trims redundant lighting parameters and demonstrates superior localization properties for illumination representation. In addition, a novel spherical transport loss is designed based on optimal transport theory which guides to regress lighting representation parameters with consideration of spatial information. Furthermore, we designed a new metric that is concise yet effective by directly evaluating the estimated illumination maps rather than rendered images. Extensive experiments show that NeedleLight achieves superior lighting estimation consistently across multiple evaluation metrics as compared with state-of-the-art methods.

1.3 Outline of the Thesis

Chapter 1 introduces the scope of this thesis and provides an overview of image synthesis task as well as issues and practical concerns involved, which essentially

motivates this research work. The main contributions are also briefly stated and discussed in this Chapter.

Chapter 2 reviews the background of image synthesis, which starts from the traditional image synthesis methods, followed by image synthesis for deep learning and further image synthesis with generative model.

Chapter 3 presents the proposed composition-based image synthesis. We investigate both traditional and learning-based methods for image composition, and demonstrate the effectiveness of the synthetic images in deep network training.

Chapter 4 presents our translation-based methods for image synthesis. We achieve the image translation in geometry domain or both geometry and appearance domains for achieving optimal image translation realism. Experiments on deep network training and user study demonstrate the superiority of our synthesized images.

Chapter 5 presents our proposed 3D-level image synthesis with realistic illuminations. We explore two approaches to achieve realistic illumination estimation, one generating illumination effects implicitly through adversarial learning and and the other recover illumination map of background scenes explicitly.

Chapter 6 summarizes the thesis and envisions the future works.

Chapter 2

Literature Review

This chapter reviews the development of image synthesis. We start from the traditional image synthesis (Section 2.1), followed by image synthesis using deep learning techniques (Section 2.2) and further adversarial image synthesis (Section 2.3).

2.1 Traditional Image Synthesis

Automated image generation using computers has been a research topic since the mid of 20th century. The research on this area is empowered by the establishment of fundamental concepts such as shading [8, 9] and ray tracing [7]. The research progress of this area further enables the wide application of image generation, ranging from computer games and advertisements to virtual reality, etc. The technical approaches can be broadly classified into two categories, namely, rendering-based image synthesis and composition-based image synthesis.

2.1.1 Rendering-based Image Synthesis

The pipeline of rendering-based image synthesis consists of two parts including content generation and rendering. Content generation denotes the construction of virtual objects, which may range from simple virtual objects to complicate virtual environments which requires to build realistic geometries, illumination and materials, etc. Numbers of rendering-based image synthesis methods employ simulated

sensors to capture visual data by moving it in virtual environment. Although numbers of tools can be adopted to build the virtual environments, most content generation processes still rely heavily on manual design. An alternative is to only generate what is currently in the view of simulated sensors [39, 40] instead of generating the full virtual environments. The scene configuration is often a mixture of online objects and objects from model libraries. Taking the synthetic data for scene parsing [40] as an example, the sidewalks, buildings and traffic lights are produced by online generation, while the models of cars, bicyclists and pedestrians are selected from model libraries.

Depending on the trade-off between accuracy and computational complexity, rendering can be performed with several different approximations. The light transport theory [41] formulate the transport of light from light sources to the camera via the medias or surfaces in scene. For surfaces, the rendering equations [42] is often adopted to describe the light transport. Image rendering is carried out by estimating the equilibrium rendering equation, which is challenging as the outgoing radiance from any points affects the incident radiance at every other point in the scene. In practice, the rendering problem can be solved using numerical integration, which can be conducted in several ways with light transport models and approximations. Rasterization and path tracing are two main classes of techniques to solve the rendering problem by using different approximations. Specially, rasterization is generally used in GPU-based rendering, while path tracing usually solves the rendering equation by Monte Carlo integration. With GPU is employed in rendering, rasterization techniques are generally significantly faster. However, significant pre-computations are usually required to achieve accurate sensor simulation and realistic scene representations. Path-tracing techniques can accurately simulate any type of light transport without pre-computations. This generality is a significant for many image rendering systems, as it is often required to generate large numbers of diverse samples.

2.1.2 Composition-based Methods

Thanks to the high computation cost of rendering-based methods, a large body of works [43] aim to synthesize images by inserting contents (i.e. foreground objects) into scene images (i.e. background images) with synthesis realism in object sizes, object perspectives (or orientations), insertion positions, environmental lighting, etc. One of the challenges of image insertion (or compositing) is that the foreground objects and background images usually come from different scenes and present different features and distributions, which often negatively effects the composition realism. The differences can be in either appearance domain such as lighting, white balance and shading or geometry domain such as viewpoint and positioning. Numbers of prior works have attempted to correct the appearance through alpha blending [44], Poisson blending [45] or recent deep learning based approaches [46, 47]. To further improve the synthesis realism, all scene properties including geometry, surface reflectance, and illumination are expected to be inferred from the scene to assist the image synthesis. For example, Karsch et al. [13] develop a semi-automatic technique that inserts objects into legacy photographs with photo-realistic perspective and illumination, in which the object size is determined according to the method proposed by Criminisi et al. [48]. Besides, Hoiem et al. [49] propose to determine the insertion geometry by dividing the image regions into vertical plane, ground plane or sky.

As scene illuminations play a critical role for the visual quality of synthesized images, a large body of works [43, 50–55] focus on reconstructing illuminations from the scene. Specially, Karsh et al. [13] recover parametric 3D lighting from a single image with user annotations for initial lighting and geometry estimation. Barron et al. [52] estimate shape, lighting and material and rely on data-driven priors to compensate for the lack of geometry information. Sengupta et al. [56] jointly estimate albedo, normals, and lighting of an indoor scene from a single image. Lombardi et al. [57] estimate illumination from an object of known shape with a low-dimensional model. Maier et al. [58] make use of additional depth information to recover spherical harmonics illumination. On the other hand, the recent learning-based works aim to estimate lighting from images by regressing representation parameters [3, 59, 60] or generating illumination maps [61, 62]. Cheng et al. [59] regress spherical harmonics (SH) parameters of global lighting with a render loss. Garon et al. [2] estimate lighting by predicting spherical harmonic (SH)

coefficients from a background image and local patch. Gardner et al. [3] estimate the positions, intensities, and colours of light sources and reconstructs illumination maps with a spherical Gaussian function. On top of it, Li et al. [63] represents illumination maps with multiple spherical Gaussian functions and regresses the corresponding Gaussian parameters for lighting estimation. Gardner et al. [61] generate illumination maps directly with a two-steps training strategy. Song et al. [62] estimate per-pixel 3D geometry and uses a convolutional network to predict unobserved contents in the environment map. Legendre et al. [64] regress HDR lighting from LDR images by comparing the rendered sphere with predicted illumination to the ground truth. Srinivasan et al. [65] estimate a 3D volumetric RGB model of a scene and uses standard volume rendering to estimate incident illuminations.

2.2 Image Synthesis for Data Augmentation

Recent progress in computer vision and deep learning has been driven by the large amount of training data. Unfortunately, creating large datasets with accurate labels has been extremely costly due to the amounts of human efforts required. Thus, synthetic data has emerged as a cheap yet efficient method for data augmentation in network training and model evaluations.

Several works explored to utilize 3D synthetic data to tackle 2D computer vision problems such as face recognition, object detection [66] and scene understanding [67]. Computer vision researchers previously leveraged 3D simulations to model hand appearance [68], human shapes [69], and inverse graphics problem [70]. On the other hand, these methods suffer from several limitations by requiring manually tuning and controlled virtual environments In addition, the lack of photo-realism introduces a significant domain discrepancy between synthetic and real data, which might cause the synthetic data less effective in network training.

The recent progress in computer graphics improves the realism of synthesized images. The modern graphics platforms also enable the widespread use of synthetic data that are generated with less constraint. The work by Vazquez et al. [71] is early attempts to use synthetic data for training and explore if the appearance models of pedestrians can be learnt in a virtual world. Gaidon et al. [72] propose

an efficient real-to-virtual cloning method, and validate the method by constructing a dataset that includes automatically generated labels for tracking, detection, scene segmentation, etc. In the scene of video surveillance, Taylor et al. [73] propose a virtual simulation platform for model design and evaluations. Other works explore the application of 3D models for object pose estimation [74] and detection [75]. Besides, Richter et al. [76] present an approach to create images rapidly with pixel-wise annotation from modern computer games; Aldrian et al. [14] propose an inverse rendering method to generate 3D face structures; Yildirim et al. [16] use the CNN features trained on synthetic faces to regress face pose parameters. Rendered depth maps have been used prominently in model training for hand tracking and pose estimation [77, 78]. 3D models are also widely used for learning feature representations for object detection and pose estimation [75, 79, 80]. Renderings of 3D models have been applied to train neural networks for optical flow estimation [15] and shape-from-shading algorithms [81]. Renderings of full synthetic environments have been proposed for training deep models for scene flow estimation [82].

Several works [83] utilize synthetic data for evaluation purposes and most works focus on low-level vision tasks. For example, Butler et al. [84] propose a synthetic benchmark for the estimation of optical flow. Chen et al. [85] is typical example for basic building blocks of autonomous driving. These approaches treat synthetic data as an easy approach to obtain ground truth.

Among various synthetic data for model training or evaluation, the synthesis of scene text images is emerged thanks to the excellent performance of synthetic data for the training of scene text detection and recognition models. Specially, Jaderberg et al. [6] create a word generator and use the synthetic images to train text recognition networks. Gupta et al. [5] develop a fast and scalable engine to generate synthetic images of texts in scenes. Long and Yao [86] introduce an efficient image synthesis method that renders realistic scene text images via a 3D graphics engine. Yang et al. [87] propose a learning-based data-driven text synthesis engine for scene text detection task. Another typical application of image synthesis is to boost the performance of person re-identification. For example, Barbosa et al. [88] propose a synthetic instance dataset SOMAset, which is created by photo-realistic human body generation software. Bak et al. [89] construct a SyRI dataset by using 100 virtual humans illuminated with multiple HDR environment maps. In

addition, Sun et al. [90] introduce a large-scale synthetic data engine named PersonX. The aforementioned methods are neither intensively diversified nor editable and extendable by the public. Xiang et al. [91] recently address this constraint by proposing GPR800 synthetic dataset which has configurable backgrounds, weathers, illuminations and diversified identities, and can be extended for constructing large-scale labeled datasets.

2.3 Generative Models for Image Synthesis

With the advance of deep neural networks in recent years, generative networks have attracted increasing attention due to their promising synthesis performance. The recent studies of the generative networks can be broadly categorized into three groups including unsupervised fundamental networks, auto-encoder networks, generative adversarial networks.

2.3.1 Unsupervised Fundamental Networks

Unsupervised fundamental networks were explored extensively for texture synthesis while they often suffer from blurry textures in generated images.

As a type of Boltzmann Machines (BM), Restricted Boltzmann Machine (RBM) [92] has been applied to various image generation tasks with improved computational costs.

By utilizing the generative decoders with Gibbs sampling, RBM and Deep Boltzmann Machine(DBM) [93] can reconstruct input image from the latent representation features. Markov chain Monte Carlo (MCMC) method [94] has also been studied to generate fair samples through the introduction of random dice. Besides, Deep Belief Networks (DBN) [95] with the stacked version of RBM adopt a multi-layer structure to provide high-level feature representations [96].

2.3.2 Auto-Encoder Networks

Auto-Encoder (AE) [97] combines an encoder and a decoder in a single network for image generation. The decoder serves as a generator to generate the data representation from the features encoded by the encoder. Various auto-encoder models have been explored in previous work. For example, Denoising Auto-encoder (DAE) model [98] is introduced for representation learning with corrupted data as input. Bengio et al. [99] propose Stacked Auto-encoder (SAE) to train auto-encoder with a layer-wise manner. Vincent et al. propose Stacked Denoising Auto-encoder (SDAE) [100] where DAE is stacked within a single model. Furthermore, Kingma et al. [101] introduce a deep generative model based on variational methods to achieve image generation. Generalized DAE (GDAE) [102] is proposed to generate images by learning a Markov-chain transition operator, which is further extended by Generative Stochastic Network (GSN) [103]. Adversarial Auto-encoder (AAE) [104] is another auto-encoder by replacing KL-divergence with generative adversarial term. Variational Auto-encoder (VAE) [105] introduces a variational component to deal with the gradient descent and over-fitting problem.

Due to the limited disentangled performance of VAE, β Variational Autoencoder (VAE) [106] is proposed with additional hyper-parameter to learn disentangled representations by balancing latent code capacity and independence generative factors. Denoising Variational Autoencoder (DVAE) [107] combines VAE and DAE to build robust models by injecting noises in both the input level and a stochastic hidden layer. Tolstikhin et al. propose Wasserstein Autoencoder (WAE) [108] which minimizes the loss based on Wasserstein distance. Deep Recurrent Attentive Writer (DRAW) [109] is another AE-based recurrent model with attention mechanism. Multiple auto-encoders were incorporated with the recurrent structure and the attention mechanism enable the network to focus on parts of the images.

2.3.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [17] have achieved great success in generating new images from either existing images or random noises. Instead of manually selecting features and parameters, GAN generators learn an optimal mapping from certain input to certain synthesized images, while GAN discriminators differentiate

the synthesized images from natural ones via adversarial learning. Based on the generation manner, GANs-based image synthesis can be broadly classified into several categories including generation from noise, generation from conditional inputs, and generation by composition.

Numbers of GAN-based image synthesis methods have explored image generation from noise. For example, [18] introduces Laplacian pyramids that improve the quality of GAN-synthesized images greatly. Other networks attempt to synthesize realistic images by stacking a pair of generators [110, 111], leaning more reasonable potential features [20], exploring new training approach [21], visualizing and understanding GANs at the unit, object and scene levels [112], etc.

To achieve effective control of the synthesized images, adversarial image generation from conditional inputs has drawn increasing attention in recent years. The conditional inputs could be semantic segmentation, key-points, edge map, another image, etc. For instance, CycleGAN [113] proposes a cycle-consistent adversarial network for realistic image-to-image translation, and so do a number of related GANs [22, 25, 27, 114–117].

Besides image generation from noise and conditional inputs, some works also explore to learn image composition in an adversarial manner. LR-GAN [118] explores to generate image background and foregrounds separately, followed by a contextually relevant manner to stitch the foregrounds on backgrounds. Lee et al. [119] propose an end-to-end trainable network for inserting an object instance mask of a specified class into the semantic label map of an image. GP-GAN [120] synthesizes high resolution images by leveraging Poisson blending [45]. Zhu et al. [121] propose a model to distinguish natural photographs from automatically generated composite images. Lin et al. [1] present a spatial transformer GAN (ST-GAN) by inserting STNs into the generator, Azadi et al. [122] describe a Compositional GAN that introduces a self-consistent composition-decomposition network, Chen et al. [123] propose a generative adversarial network (GAN) architecture for automatic image compositing with consideration of geometric, color and boundary consistency.

Chapter 3

Composition-based Image Synthesis

3.1 Introduction

Image synthesis has been attracting increasing attention as a means of generating novel images and creating annotated images for training deep neural networks (DNNs) models. Automated generation of realistic and novel images has a wide range of applications in various design tasks for artworks, advertisement posters, short movies, etc. Additionally, fast collection of large amounts of annotated training images has become the bottleneck for effective and efficient development and deployment of DNNs in various computer vision tasks. The current practice relies heavily on manual annotations, ranging from in-house annotation of small amounts of images to crowd-sourcing based annotation of large amounts of images. On the other hand, the manual annotation approach is usually expensive, time-consuming, prone to human errors and difficult to scale while data are collected under different conditions or within different environments.

Three approaches have been investigated to cope with the image annotation challenge in DNN training to the best of our knowledge. The first approach is probably the easiest and most widely adopted which augments training images by various label-preserving geometric transformations such as translation, rotation and flipping, as well as different intensity alternation operations such as blurring and histogram equalization [124]. The second approach is image synthesis based which

has been widely investigated in area of computer graphics for the purpose of education, design simulation, advertising, entertainment, etc. [125]. It creates new images by modelling the physical behaviors of light and energy in combination of different rendering techniques, aiming for realistic embedding objects of interest (OOI) into a set of "background images". To make the synthesized images useful for DNN training, the OOI should be embedded in the way that it looks as natural as possible. At the same time, sufficient variations should be included to ensure that the learned representation is sufficiently diverse to capture most possible OOI appearances in real scenes. The third approach is machine learning based which employs various semi-supervised and unsupervised learning techniques to create more annotated training images. For example, bootstrapping has been studied which combines the traditional self-training and co-training with the recent DNN training to search for more training samples (with pseudo labels or annotations) from a large number of unannotated images [126, 127].

In recent years, unsupervised DNN models such as Generative Adversarial Networks (GAN) [17] have also been exploited to generate more annotated training images for DNN training [128]. However, most existing works do not fully consider semantic coherence, apt embedding locations and appearance of embedded objects which are critically important while applying the synthesized images to train DNN models. In addition, most existing GANs were designed to achieve synthesis realism either from geometry space or appearance space but few in both. Consequentially, most GAN-synthesized images have little contribution (many even harmful) when they are used in training deep network models. Particularly, direct image generation still faces difficulties in generating high-resolution images due to the limited network capacity. GAN-based image composition is capable of producing high-resolution images [1, 122] by placing foreground objects into background images. But most GAN-based image composition techniques focus on geometric realism only (e.g. object alignment with contextual background) which often produce various artifacts due to appearance conflicts between the foreground objects and the background images. As a comparison, GAN-based image-to-image translation aims for appearance realism by learning the style of images of the target domain whereas the geometric realism is largely ignored.

With a foreground object and a background image for composition, a direct combination tends to introduce unrealistic artifacts in the composed image with various

inconsistency in colors, illuminations, texture, etc. Several methods have been proposed for harmonizing the foreground object and the background image, including DNN-based image translation [46], local statistic transfer [47], etc. On the other hand, existing methods typically simplify the harmonization by solving the image consistency within the foreground object region only but neglecting that realistic image composition goes far beyond that. Specifically, a typical visual effect that is often associated with objects in scenes is their projected shadows due to the 3D nature of the world. While embedding a foreground object into a background image for image composition, the shadow effects of the foreground object are critical which can significantly improve the realism of the composed image if handled properly. Due to various constraints in 2D images, such as complicated illuminations, losing of 3D information, etc., traditional computer vision methods are highly susceptible to failure for the task of high-fidelity shadow generation.

To overcome the problem of limited training data, we designed a novel image synthesis technique [129] [1] that aims to create a large amount of annotated scene text images for training accurate and robust scene text detection and recognition models as presented in section 3.3. To achieve realistic image composition in both geometry and appearance domains, we designed an innovative Spatial Fusion GAN (SF-GAN) [33] [2] that achieves synthesis realism in both geometry and appearance spaces concurrently as presented in section 3.4. To achieve image composition with realistic illuminations and shadows, we design a novel Adversarial Image Composition Net (AIC-Net) [34] [3] that aims to achieve realistic image composition with adaptive shadow effects as presented in section 3.5.

3.2 Related Works

Image Composition

Photo-realistic image composition [1, 130] has been explored extensively which synthesizes new images by placing foreground objects into some existing background image. The target is to achieve composition realism by controlling the object size, orientation, and blending between foreground objects and background images. For

[1] The work has been published in ECCV 2018
[2] The work has been published in CVPR 2019
[3] The work has been published in ACCV 2020

example, [13] develops a semi-automatic technique that inserts objects into legacy photographs with photorealistic lighting and perspective. In recent years, image synthesis has been investigated as a data augmentation approach for training accurate and robust DNN models when only a limited number of annotated images are available. For example, Dosovitskiy et al. [131] use synthetic floating chair images to train optical flow networks. Yildirim et al. [16] use the CNN features trained on synthetic faces to regress face pose parameters. Gupta el al. [5] and Zhan et al. [132] develop a fast and scalable engine to generate text images for scene text detection and recognition [133]. [5, 129, 134] investigate synthesis of scene text images for training better scene text detection and recognition models. They achieve synthesis realism by controlling a series of parameters such as text locations within the background image, geometric transformation of the foreground texts, blending between the foreground text and background image, etc. Other image composition systems have also been reported for DNN training [135], composition harmonization [46, 47], image inpainting [136, 137], etc.

Image Harmonization

Image harmonization deals with the inconsistency between a foreground region and a background image in composition. Several methods have been proposed to adapt the appearance of the foreground object for image harmonization. For example, alpha blending [138] is adopted to make the transition as smooth as possible so that artifacts can be hidden/removed within the composed images (this approach tends to blur fine details in the foreground object and background images instead). In addition, gradient-based techniques such as Poisson blending [45] can edit image gradients and adjust the inconsistency in color and illumination to achieve seamlessly blending. Sunkavalli et al. [139] match contrast, texture, and blur by manipulating the scales of a pyramid decomposition of images. Xue et al. [140] study image statistics that determine the realism of the composite images. Tao et al. [141] propose a gradient domain to achieve composition which can prevent color bleeding without changing the boundary location.

Numbers of learning-based image harmonization methods have also proposed in recent years. For example, Tsai et al. [142] utilize visual semantics to guide the process of replacement to generate realistic, artifact-free images with diverse styles. Luan et al. [47] achieve both spatial and inter-scale statistical consistency by carefully determining the local statistics to be transferred. Tsai et al. [46] propose

an end-to-end deep convolutional neural network for image harmonization, which can capture both the context and semantic information of the composite images during harmonization. Chen et al. [123] combine a transformation network, a refinement network, a segmentation network and a pair of discriminator networks to achieve image composition via adversarial learning. Besides image harmonization, a few works [121, 143] focus on the realism assessment of the composed images. For example, Cong et al. [144] handle image harmonization through domain verification discriminator with the insight that the foreground needs to be translated to the same domain as background.

Generative Adversarial Networks

GANs [17] have achieved great success in image generation from either existing images or random noises [145]. Instead of manually selecting features and parameters, GAN generators learn an optimal mapping from random noise or existing images to synthesized images, while GAN discriminators differentiate the synthesized images from real ones via adversarial learning. Several GAN-based image synthesis methods have been reported in recent years. [18] introduces Laplacian pyramids that improve the quality of GAN-synthesized images greatly. [119] proposes an end-to-end trainable network for inserting an object instance mask into the semantic label map of an image. Other systems attempt to learning more reasonable potential features [20], exploring new training approach [21], visualizing GANs at the unit, object and scene level [112], etc. To expand the applicability, [146] generate shadows for augmented reality by modeling the mapping relation between the virtual object shadow and the real-world.

Besides generation in pixel domain, some works also explore to include geometry translation in image generation. Specially, LR-GAN [118] generates new images by applying additional spatial transformation networks (STNs) to factorize shape variations. GP-GAN [120] composes high-resolution images by using Poisson blending [45]. A few GANs have been reported in recent years for geometric realism, e.g., [1] presents a spatial transformer GAN (ST-GAN) by embedding STNs in the generator for geometric realism, [122] designs Compositional GAN that employs a self-consistent composition-decomposition network. Existing GANs synthesize images in either geometry space (e.g. ST-GAN) or appearance space (e.g. Cycle-GAN) but few in both spaces. In addition, the GAN-synthesized images are usually not

suitable for training deep network models due to the lack of annotation or synthesis realism.

Guided Filter

Guided Filters [147, 148] use one image as guidance for filtering another image which has shown superior performance in detail-preservative filtering. The filtering output is a linear transform of the guidance image by considering its structures, where the guidance image can be the input image itself or another different image. Guided filtering has been used in various computer vision tasks, e.g., [149] uses it for weighted averaging and image fusion, [150] uses a rolling guidance for fully-controlled detail smoothing in an iterative manner, [151] uses a fast guided filter for efficient image super-solution, [152] uses guided filters for high-quality depth map restoration, [153] uses guided filtering for tolerance to heavy noises and structure inconsistency, and [154] puts Guided filtering as a nonconvex optimization problem and proposes solutions via majorize-minimization [155]. Most GANs for image-to-image-translation can synthesize high-resolution images but the appearance transfer often suppresses image details such as edges and texture. How to keep the details of the original image while learning the appearance of the target remain an active research area.

3.3 Verisimilar Image Synthesis for Network Training

We propose a novel image synthesis technique [129] [4] that aims to create a large amount of annotated scene text images for training accurate and robust scene text detection and recognition models. The proposed technique consists of three innovative designs as listed:

1. It enables "semantic coherent" image synthesis by embedding texts at semantically sensible regions within the background image as illustrated in Fig. 3.1, e.g. scene texts tend to appear over the wall or table surface instead of the food or plant leaves. We achieve the semantic coherence by leveraging the

[4] The work has been published in ECCV 2018

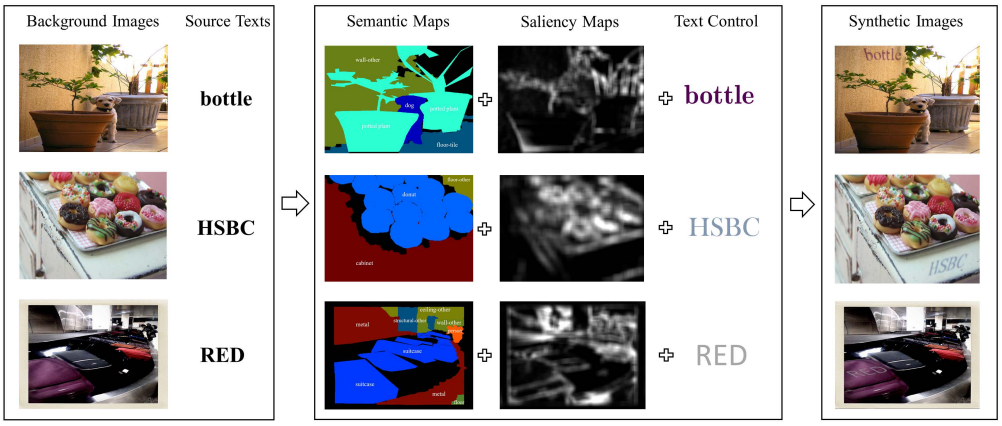


FIGURE 3.1: The proposed scene text image synthesis technique: Given background images and source texts to be embedded into the background images as shown in the left-side box, a semantic map and a saliency map are first determined which are then combined to identify semantically sensible and apt locations for text embedding. The color, brightness, and orientation of the source texts are further determined adaptively according to the color, brightness, and contextual structures around the embedding locations within the background image. Pictures in the right-side box show scene text images that are synthesized by the proposed technique.

semantic annotations objects and image regions that have been created and are readily available in the semantic segmentation research

2. It exploits visual saliency to determine the embedding locations within each semantic coherent region as illustrated in Fig. 3.1. Specifically, texts are usually placed at homogeneous regions in scenes for better visibility and this can be perfectly captured using visual saliency. The exploitation of saliency guidance helps to synthesize more natural-looking scene text images

3. It designs a novel scene text appearance model that determines the color and brightness of source texts by learning from the feature of real scene text images adaptively. This is achieved by leveraging the similarity between the neighboring background of texts in scene images and the embedding locations within the background images

The proposed scene text image synthesis technique starts with two types of inputs including "Background Images" and "Source Texts" as illustrated in columns 1 and 2 in Fig. 3.1. Given background images, the regions for text embedding can be determined by combining their "Semantic Maps" and "Saliency Maps" as illustrated in columns 3-4 in Fig. 3.1, where the "Semantic Maps" are available as



(a) (b) (c) (d)

FIGURE 3.2: Without semantic coherence (SC) as illustrated in (b), texts may be embedded at arbitrary regions such as sky and the head of sheep which are rarely spotted in scenes as illustrated in (c). SC helps to embed texts at semantically sensible regions as illustrated in (d).

ground truth in the semantic image segmentation research and the "Saliency Maps" can be determined using existing saliency models. The color and brightness of source texts can then be estimated adaptively according to the color and brightness of the determined text embedding regions as illustrated in column 5 in Fig. 3.1. Finally, "Synthesized Images" are produced by placing the rendered texts at the embedding locations as illustrated in column 6 in Fig. 3.1.

3.3.1 Semantic Coherence

Semantic coherence (SC) refers to the target that texts should be embedded at semantically sensible regions within the background images. For example, texts should be placed over the fence boards instead of sky or sheep head where texts are rarely spotted in real scenes as illustrated in Fig. 3.2. The SC thus helps to create more semantically sensible foreground-background pairing which is very important to the visual representations as well as object detection and recognition models that are learned/trained by using the synthesized images. To the best of our knowledge, SC is largely neglected in earlier works that synthesize images for better deep network model training, e.g. the recent work [5] that deals with a similar scene text image synthesis problem.

We achieve the semantic coherence by exploiting large amounts of semantic annotations objects and image regions that have been created in the semantic image segmentation research. Particularly, a number of semantic image segmentation datasets [156] have been created each of which comes with a set of "ground truth" images that have been "semantically" annotated. The ground truth annotation divides an image into numbers of objects or regions at the pixel level where each

object or region has a specific semantic annotation such as "cloud", "tree", "person", "sheep", etc. as illustrated in Fig. 3.2.

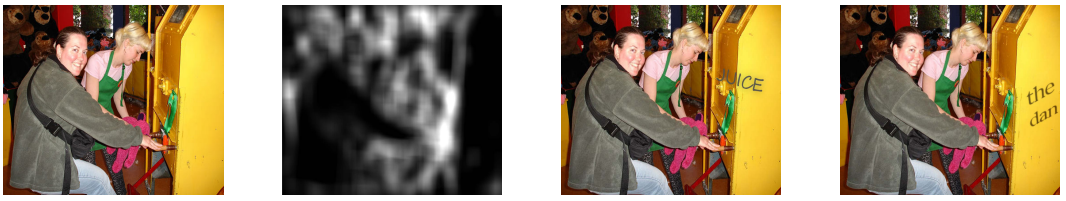
To exploit SC for semantically sensible image synthesis, all available semantic annotations within the semantic segmentation datasets [156] are first classified into two lists where one list consists of objects or image regions that are semantically sensible for text embedding and the other consists of objects or image regions not semantically sensible for text embedding. Given some source texts for embedding and background images with region semantics, the image regions that are suitable for text embedding can thus be determined by checking through the pre-defined list of region semantics.

3.3.2 Saliency Guidance

Not every location within the semantically coherent objects or image regions are suitable for scene text embedding. For example, it is more suitable to embed scene texts over the surface of the yellow-color machine instead of across the two neighboring surfaces as illustrated in (c) and (d) of Fig. 3.3. Certain mechanisms are needed to further determine the exact scene text embedding locations within semantically coherent objects or image regions.

We exploit the human visual attention and scene text placement principle to determine the exact scene text embedding locations. To attract the human attention and eyeballs, scene texts are usually placed around homogeneous regions such as signboards to create good contrast and visibility. With such observations, we make use of visual saliency as a guidance to determine the exact scene text embedding locations. Particularly, homogeneous regions usually have lower saliency as compared with those highly contrasted and cluttered. Scene texts can thus be place at locations that have low saliency within the semantically coherent objects or image regions as described in the last subsection.

Quite a number of saliency models have been reported in the literature [157]. We adopt the saliency model in [158] due to its good capture of local and global contrast. Given an image, the saliency model computes a saliency map as illustrated in Fig. 3.3 (b), where homogeneous image regions usually have lower saliency. The



(a) (b) (c) (d)

FIGURE 3.3: Without saliency guidance (SG) as illustrated in (b), texts may be embedded across the object boundary as illustrated in (c) which are rarely spotted in scenes. SG thus helps to embed texts at right locations within the semantically sensible regions as illustrated in (d).

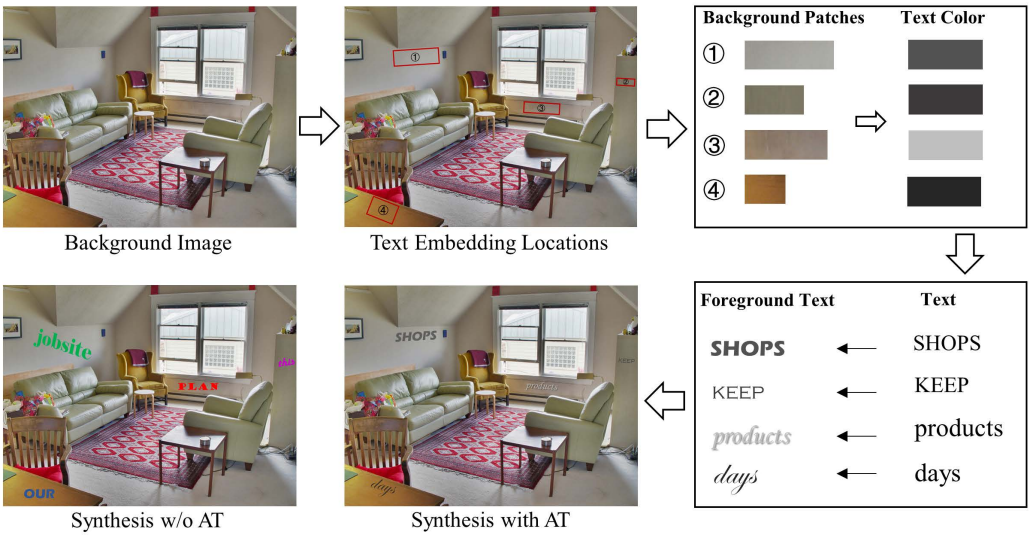


FIGURE 3.4: Adaptive text appearance (ATA): The color and brightness of source texts are determined adaptively according to the color and brightness of the background image around the embedding locations as illustrated. The orientations of source texts are also adaptively determined according to the orientation of the contextual structures around the embedding locations. The ATA thus helps to produce more verisimilar text appearance as compared random setting of text color, brightness, and orientation.

locations that are suitable for text embedding can thus be determined by thresholding the computed saliency map. In our implemented system, a global threshold is used which is simply estimated by the mean of the computed saliency map. As Fig. 3.3 shows, the saliency guidance helps to embed texts at right locations within the semantically sensible regions. The use of saliency guidance further helps to improve the verisimilitude of the synthesized images as well as the learned visual representation of detection and recognition models.

3.3.3 Adaptive Text Appearance

Visual contrast as observed by low-level edges and corners is crucial feature while training object detection and recognition models. Texts in scenes are usually presented by linear strokes of different sizes and orientations which are rich in contrast-induced edges and corners. Effective control of the contrast between source texts and background images is thus very important to the usefulness of the synthesized images while applying them to train scene text detection and recognition models.

We design an adaptive contrast technique that controls the color and brightness of source texts according to what they look like in real scenes. The idea is to search for scene text image patches (readily available in a large amount of scene text annotations within existing datasets) whose background has similar color and brightness to the determined background regions. The color and brightness of the source texts can then be determined by referring to the color and brightness of text pixels within the searched scene text image patches.

The scene text image patches are derived from the scene text annotations as readily available in existing datasets such as ICDAR2013 [159]. For each text annotation, a HoG (histogram of oriented gradient) feature H_b is first built by using the background region surrounding the text annotation under study. The mean and standard deviation of the color and brightness of the text pixels within the annotation box are also determined in the Lab color space, as denoted by (μ_L, σ_L), (μ_a, σ_a) and (μ_b, σ_b). The background HoG H_b and the text color and brightness statistics (μ_L, σ_L), (μ_a, σ_a) and (μ_b, σ_b) of a large amount of scene text patches thus form a list of pairs as follows:

$$P = \left\{ H_{b_1} : (\mu_{L_1}, \sigma_{L_1}, \mu_{a_1}, \sigma_{a_1}, \mu_{b_1}, \sigma_{b_1}), \cdots H_{b_i} : (\mu_{L_i}, \sigma_{L_i}, \mu_{a_i}, \sigma_{a_i}, \mu_{b_i}, \sigma_{b_i}), \cdots \right\} \quad (3.1)$$

The H_b in Eq. 1 will be used as the index of the annotated scene text image patch, and (μ_L, σ_L), (μ_a, σ_a) and (μ_b, σ_b) will be used as a guidance to set the color and brightness of the source text. For each determined background patch (suitable for text embedding) as illustrated in Fig. 3.4, its HoG feature H_s can be extracted and the scene text image patch that has the most similar background can thus be determined based on the similarity between H_s and H_b. The color and brightness of the source text can thus be determined by taking the corresponding (μ_L, μ_a, μ_b) plus random variations around (σ_L, σ_a, σ_b).

The proposed technique also controls the orientation of the source texts adaptively according to certain contextual structures lying around the embedding locations within the background image. Particularly, certain major structures (such as the table borders and the boundary between two connected wall surfaces as illustrated in Fig. 3.4) as well as their orientation can be estimated from the image gradient. The orientation of the source texts can then be determined by aligning with the major structures detected around the scene text embedding locations as illustrated in Fig. 3.4. Beyond the text alignment, the proposed technique also controls the font of the source texts by randomly selecting from a pre-defined font list as illustrated in Fig. 3.4.

3.3.4 Experiments

We evaluate the effectiveness of the proposed image synthesis technique on a scene text detection task and a scene text recognition task. The evaluations are performed over 5 public datasets to be discussed in the following subsections.

Implementations

Scene Text Detection: We use an adapted version of EAST [160] to train all scene text detection models. EAST is a simple but powerful detection model that yields fast and accurate scene text detection in scene images. The model directly predicts words or text lines of arbitrary orientations and quadrilateral shapes in the images. It utilizes the fully convolutional network (FCN) model that directly produces words or text-line level predictions, excluding unnecessary and redundant intermediate steps. Since the implementation of the original EAST is not available, we adopt an adapted implementation that uses ResNet-152 instead of PVANET [161] as the backbone network.

Scene Text Recognition For the scene text recognition, we use the CRNN model [162] to train all scene text recognition models. The CRNN model consists of the convolutional layers, the recurrent layers and a transcription layer which integrates feature extraction, sequence modelling and transcription into a unified framework. Different from most existing recognition models, the architecture in CRNN is end-to-end trainable and can handles sequences in arbitrary lengths, involving no character segmentation. Moreover, it is not confined to any predefined lexicon and

can reach superior recognition performances in both lexicon-free and lexicon-based scene text recognition tasks.

Datasets and Evaluation Metrics: The proposed technique is evaluated over five public datasets including ICDAR 2013 [159],ICDAR 2015 [163], MSRA-TD500 [164], IIIT5K [165] and SVT[166].

ICDAR 2013 dataset is obtained from the Robust Reading Challenges 2013. It consists of 229 training images and 233 test images that capture text on sign boards, posters, etc. with word-level annotations. For recognition task, there are 848 images for training recognition models and 1095 images for recognition model evaluation. We use this dataset for both scene text detection and scene text recognition evaluations.

ICDAR 2015 is a dataset of incidental scene text and consists of 1,670 images (17,548 annotated text regions) acquired using the Google Glass. Incidental scene text refers to text that appears in the scene without the user taking any prior action in capturing. We use this dataset for the scene text detection evaluation.

MSRA-TD500 dataset consists of 500 natural images (300 for training, 200 for test), which are taken from indoor and outdoor scenes using a pocket camera. The indoor images mainly capture signs, doorplates and caution plates while the outdoor images mostly capture guide boards and billboards with complex background. We use this dataset for the scene text detection evaluation.

IIIT5K dataset consists of 2000 training images and 3000 test images that are cropped from scene texts and born-digital images. For each image, there is a 50-word lexicon and a 1000-word lexicon. All lexicons consist of a ground truth word and some randomly picked words. We use this dataset for scene text recognition evaluation only.

SVT dataset consists of 249 street view images from which 647 words images are cropped. Each word image has a 50-word lexicon. We use this dataset for scene text recognition evaluation only.

For the scene text detection task, we use the evaluation algorithm by Wolf et al. [167]. For the scene text recognition task, we perform evaluations based on the correctly recognized words (CRW) which can be calculated according to the ground truth transcription.

TABLE 3.1: Scene text detection recall (R), precision (P) and f-score (F) on the ICDAR2013, ICDAR2015 and MSRA-TD500 datasets, where "*EAST*" denotes the adapted EAST model, "Real" denotes the original training images within the respective datasets, "Synth 1K" and "Synth 10K" denote 1K and 10K synthesized images by our method.

Methods	ICDAR2013			ICDAR2015			MSRA-TD500		
	R	P	F	R	P	F	R	P	F
I2R NUS FAR [159]	73.0	66.0	69.0	-	-	-	-	-	-
TD-ICDAR [164]	-	-	-	-	-	-	52.0	53.0	50.0
NJU [163]	-	-	-	36.3	70.4	47.9	-	-	-
Kang et al. [168]	-	-	-	-	-	-	62.0	71.0	66.0
Yin et al. [169]	65.1	84.0	73.4	-	-	-	63.0	81.0	71.0
Jaderberg et al. [170]	68.0	86.7	76.2	-	-	-	-	-	-
Zhang et al. [171]	78.0	88.0	83.0	43.1	70.8	53.6	67.0	83.0	74.0
Tian et al. [172]	83.0	93.0	88.0	51.6	74.2	60.9	-	-	-
Yao et al. [173]	80.2	88.9	84.3	58.7	72.3	64.8	**76.5**	75.3	75.9
Gupta et al. [5]	76.4	**93.8**	84.2	-	-	-	-	-	-
Zhou et al. [160]	82.7	92.6	87.4	**78.3**	83.3	80.7	67.4	**87.3**	76.1
EAST (Real)	80.5	85.6	83.0	75.8	84.1	79.7	69.2	78.1	73.4
EAST (Real+Synth 1K)	83.5	89.3	86.3	76.2	85.4	80.5	70.6	80.9	75.4
EAST (Real+Synth 10K)	**85.0**	91.7	**88.3**	77.2	**87.1**	**81.9**	72.7	85.7	**78.6**

Scene Text Detection: For the scene text detection task, the proposed image synthesis technique is evaluated over three public datasets ICDAR2013, ICDAR2015 and MSRA-TD500. We synthesize images by catering to specific characteristics of training images within each dataset in term of text transcripts, text languages, text annotation methods, etc. Take the ICDAR2013 dataset as an example. The source texts are all in English and the embedding is at word level because almost all texts in the ICDAR2013 are in English and annotated at word level. For the MSRA-TD500, the source texts are instead in a mixture of English and Chinese and the embedding is at text line level because MSRA-TD500 contains both English and Chinese texts with text line level annotations. In addition, the source texts are a mixture of texts from the respective training images and publicly available corpses. The number of embedded words or text lines is limited at the maximum of 5 for each background image since we have sufficient background images with semantic segmentation.

Table 3.1 shows experimental results by using the adapted EAST (denoted by *EAST*) model. For each dataset, we train a baseline model "*EAST* (Real)" by using the original training images only as well as two augmented models "*EAST* (Real+Synth 1K)" and "*EAST* (Real+Synth 10K)" that further include 1K and 10K our synthesized images in training, respectively. As Table 3.1 shows, the scene

TABLE 3.2: Scene text detection performance on the ICDAR2013 dataset by using the adapted EAST model, where "Synth" and "Gupta" denote images synthesized by our method and Gupta et al. [5] respectively, "1K" and "10K" denote the number of synthetic images used, "Random" means embedding texts at random locations, SC, SG and ATA refer to semantic coherence, saliency guidance, and adaptive text appearance.

Training Data	Recall	Precision	F-measure
ICDAR2013 (Baseline)	80.49	85.56	82.95
ICDAR2013 + 1k Synth (Random)	81.66	84.49	83.08
ICDAR2013 + 1k Synth (SC)	82.15	86.34	84.19
ICDAR2013 + 1k Synth (SG)	82.26	87.33	84.72
ICDAR2013 + 1k Synth (ATA)	81.90	84.95	83.40
ICDAR2013 + 1k Synth (SC+SG)	82.74	89.39	85.94
ICDAR2013 + 1k Synth (SC+ATA)	82.79	87.54	85.10
ICDAR2013 + 1k Synth (SG+ATA)	82.70	88.95	85.72
ICDAR2013 + 1k Synth (SC+SG+ATA)	83.46	89.34	86.29
ICDAR2013 + 10k Synth (SC+SG+ATA)	85.02	91.74	88.25
ICDAR2013 + 1k Gupta [5]	82.81	89.01	85.80
ICDAR2013 + 10k Gupta [5]	84.93	90.74	87.74

text detection performance is improved consistently for all three datasets when synthesized images are included in training. In addition, the performance improvements become more significant when the number of synthesis images increases from 1K to 10K. In fact, the trained models outperform most state-of-the-art models when 10K synthesis images are used, and we can foresee further performance improvements when a larger amount of synthesis images are included in training. Furthermore, we observe that the performance improvements for the ICDAR2015 dataset are not as significant as the other two datasets. The major reason is that the ICDAR2015 images are videos frames as captured by Google glass cameras many of which suffer from motion and/or out-of-focus blur, whereas our image synthesis pipeline does not include image blurring function. We conjecture that the scene text detection models will perform better for the ICDAR2015 dataset if we incorporate the image blurring into the image synthesis pipeline.

In particular, a f-score of 83.0 is obtained for the ICDAR2013 dataset when the model is trained using the original training images. The f-score is improved to 86.2 when 1K synthetic images are included, and further to 88.3 when 10K synthetic images are included in training. Similar improvements are observed for the ICDAR2015 dataset, where the f-score is improved from the baseline 79.7 to 80.5 and

81.9 when 1K and 10K synthetic images are included in training. For the MSRA-TD500, a f-score of 73.4 is obtained when only the original 300 training images are used in model training. The f-score is improved to 75.4 and 78.6 respectively, when 1K and 10K synthetic images are included in training. This further verifies the effectiveness of the synthesized scene text images that are produced by our proposed technique.

We also perform ablation study of the three proposed image synthesis designs including semantic coherence (SC), saliency guidance (SG) and adaptive text appearance (ATA). Table 3.2 shows the experimental results over the ICDAR2013 dataset. As Table 3.2 shows, the inclusion of synthesized images (including random embedding in "ICDAR2013 + 1k Synth (Random)") consistently improves the scene text detection performance as compared with the baseline model "ICDAR2013 (Baseline)" that is trained by using the original training images only. In addition, the inclusion of either one of our three designs help to improve the scene text detection performance beyond the random embedding, where SG improves the most as followed by SC and AC. When all three designs are included, the f-score reaches 86.26 which is much higher than 83.09 by random embedding. Furthermore, the f-score reaches 88.25 when 10K synthesized images are included in training. We also compared our synthesized images with those created by Gupta et al. [5] as shown in Table 3.2, where the scene text detection models using our synthesized training images show superior performance consistently.

Scene Text Recognition: For the scene text recognition task, the proposed image synthesis technique is evaluated over three public datasets ICDAR2013, IIIT5K and SVT as shown in Table 3.3 where the CRNN is used as the recognition model. The baseline model "CRNN (Real)" is trained by combining all annotated word images within the training images of the three datasets. As Table 3.3 shows, the baseline recognition accuracy is very low because the three datasets contain around 3100 images only. As a comparison, the recognition model "CRNN (Real+Ours 5M)" achieves state-of-the-art performance, where the 5 million images are directly cropped from our synthesized scene text images as described in the last subsection. The significant recognition accuracy improvement demonstrates the effectiveness of the proposed scene text image synthesis technique.

In particular, the correctly recognized words (CRW) increases to 87.1% for the

TABLE 3.3: Scene text recognition performance over the ICDAR2013, IIIT5K and SVT datasets, where "50" and "1K" in the second row denote the lexicon size and "None" means no lexicon used. "Real" denote the original training images, "Ours 5M", "Jaderberg 5M" and "Gupta 5M" denote the 5 million images synthesized by our method, Jaderberg et al. [6] and Gupta et al. [5] respectively.

Methods	ICDAR2013	IIIT5K			SVT	
	None	50	1k	None	50	None
Mishra et al. [174]	-	64.1	57.5	-	73.2	-
Rodrguez et al. [175]	-	76.1	57.4	-	70.0	-
Yao et al. [176]	-	80.2	69.3	-	75.9	-
Almaz´an et al. [177]	-	91.2	82.1	-	74.3	-
Gordo [178]	-	93.3	86.6	-	91.8	-
Jaderberg et al. [134]	81.8	95.5	89.6	-	93.2	71.7
Shi et al. [162]	86.7	97.6	94.4	78.2	96.4	80.8
Bissacco et al. [179]	87.6	-	-	-	90.4	78.0
Shi et al. [180]	**88.6**	96.2	93.8	**81.9**	95.5	**81.9**
CRNN (Real)	31.2	64.4	54.4	38.7	62.1	35.5
CRNN (Real+Jaderberg 5M [6])	85.6	97.1	93.2	77.1	95.6	79.9
CRNN (Real+Gupta 5M [5])	86.4	96.7	92.4	76.0	95.3	79.2
CRNN (Real+Ours 5M)	87.1	**98.1**	**95.3**	79.3	**96.7**	81.5

ICDAR2013 dataset (without using lexicon) when 5 million synthetic images (synthesized by our proposed method) are included in training. This CRW is significantly higher than the baseline 31.2% when only the original 3100 images are used in training. For the IIIT5K, the CRW is increased to 79.3% (no lexicon) when the same 5 million images are included in training. The CRW is further improved to 95.3% and 98.1%, respectively, when the lexicon size is 1K and 50. Similar CRW improvements are also observed on the SVT dataset as shown in Table 3.3.

We also benchmark our synthesized images with those created by Jaderberg et al. [6] and Gupta et al. [5]. In particular, we take the same amounts of synthesized images (5 million) and train the scene text recognition model "CRNN (Real+Jaderberg 5M [6])" and "CRNN (Real+Gupta 5M [5])" by using the same CRNN network. As Table 3.3 shows, the model trained by using our synthesized images outperforms the models trained by using the "Jaderberg 5M" and "Gupta 5M" across all three datasets. Note that the model by Shi et al. [162] achieves similar accuracy as the "CRNN (Real+Ours 5M)", but it uses 8 million synthesized images as created by Jaderberg et al. [6].

The superior scene text recognition accuracy as well as the significant improvement in the scene text detection task as described in the last subsection is largely due to the three novel image synthesis designs which help to generate verisimilar scene



FIGURE 3.5: Several sample images from our synthesis dataset that show how the pro-posed semantic coherence, saliency guidance and adaptive text appearance work together for verisimilar text embedding in scene images automatically.

text images as illustrated in Fig. 3.5. As Fig. 3.5 shows, the proposed scene text image synthesis technique is capable of embedding source texts at semantically sensible and apt locations within the background image. At the same time, it is also capable of setting the color, brightness and orientation of the embedded texts adaptively according to the color, brightness, and contextual structures around the embedding locations within the background image.

3.4 Spatial Fusion GAN for Image Synthesis

We propose an innovative Spatial Fusion GAN (SF-GAN) [33] [5] that achieves synthesis realism in both geometry and appearance spaces concurrently, a very challenging task in image synthesis due to a wide spectrum of conflicts between the foreground objects and background images with respect to relative scaling, spatial alignment, appearance style, etc. The SF-GAN address these challenges by designing a geometry synthesizer and an appearance synthesizer. The geometry synthesizer learns the local geometry of background images with which the

[5]The work has been published in CVPR 2019



FIGURE 3.6: The proposed SF-GAN is capable of synthesizing realistic images concurrently in geometric and appearance spaces. Rows 1 and 2 show a few synthesized scene text images and row 3 shows a few hat-wearing and glass-wearing images where the foreground texts, glasses and hats as highlighted by red-color boxes are composed with the background scene and face images harmoniously.

foreground objects can be transformed and placed into the background images unanimously. A discriminator is employed to train a spatial transformation network, targeting to produce transformed images that can mislead the discriminator. The appearance synthesizer learns to adjust the color, brightness and styles of the foreground objects for proper matching with the background images with minimum conflicts. A guided filter is introduced to compensate the detail loss that happens in most appearance-transfer GANs. The geometry synthesizer and appearance synthesizer are inter-connected as mutual references which can be trained end-to-end with little supervision. Fig. 3.6 illustrates some images synthesized by the proposed SF-GAN.

The contributions of this work are threefold. First, it designs an innovative SF-GAN, an end-to-end trainable network that concurrently achieves synthesis realism in both geometry and appearance spaces. To the best of our knowledge, this is the first GAN that can achieve synthesis realism in geometry and appearance spaces concurrently. Second, it designs a fusion network that introduces guided filters for detail preserving for appearance realism, whereas most image-to-image-translation GANs tend to lose details while performing appearance transfer. Third, it investigates and demonstrates the effectiveness of GAN-synthesized images in training deep recognition models, a very important issue that was largely neglected

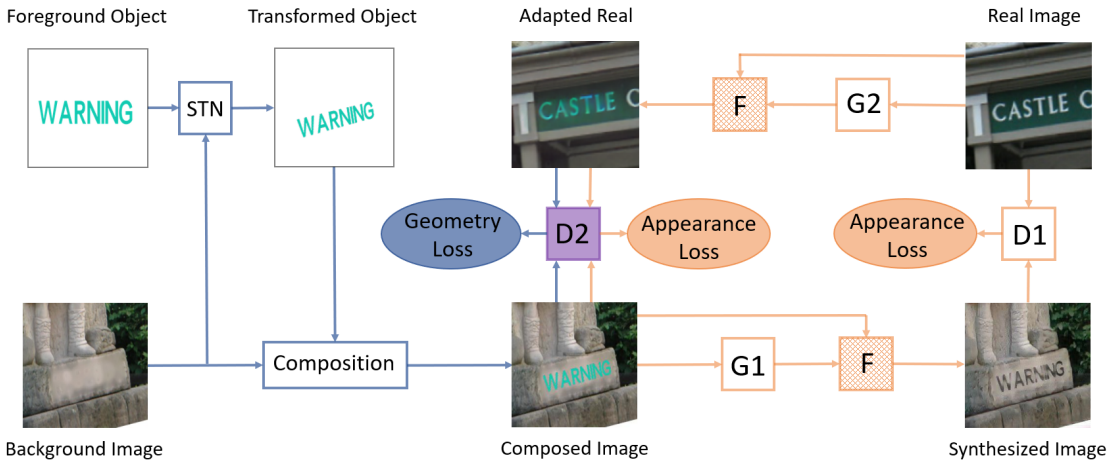


FIGURE 3.7: The structure of the proposed SF-GAN: The geometry synthesizer is highlighted by blue-color lines and boxes on the left and the appearance synthesizer is highlighted by orange-color lines and boxes on the right. *STN* denotes spatial transformation network, *F* denotes guided filters, *G1*, *G2*, *D1* and *D2* denote the generators and discriminators. For clarity, cycle loss and identity loss are not included.

in most existing GANs (except a few GANs for domain adaptation [25, 113, 181, 182]).

The proposed SF-GAN consists of a geometry synthesizer and an appearance synthesizer, and the whole network is end-to-end trainable as illustrated in Fig. 3.7. Detailed network structure and training strategy will be introduced in the following subsections.

3.4.1 Geometry Synthesizer

The geometry synthesizer has a local GAN structure as highlighted by blue-color lines and boxes on the left of Fig. 3.7. It consists of a spatial transform network (STN), a composition module and a discriminator. The STN consists of an estimation network and a transformation matrix which has N parameters that control the geometric transformation of the foreground object.

The foreground object and background image are concatenated to act as the input of the STN, where the estimation network will predict a transformation matrix to transform the foreground object. The transformation can be affine, homography, or thin plate spline [183] (We use thin plate spline for the scene text synthesis task and homography for the portrait wearing task). Each pixel in the transformed

image is computed by applying a sampling kernel centered at a particular location in the original image. With pixels in the original and transformed images denoted by $P^s = (p_1^s, p_2^s, \ldots, p_N^s)$ and $P^t = (p_1^t, p_2^t, \ldots, p_N^t)$, we use a transformation matrix H to perform pixel-wise transformation as follows:

$$\begin{bmatrix} x_i^t \\ y_i^t \\ 1 \end{bmatrix} = H \begin{bmatrix} x_i^s \\ y_i^s \\ 1 \end{bmatrix} \tag{3.2}$$

where $p_i^s = (x_i^s, y_i^s)$ and $p_j^t = (x_i^t, y_i^t)$ denote the coordinates of the i-th pixel within the original and transformed image, respectively.

The transformed foreground object can thus be placed into the background image to form an initially composed image (*Composed Image* in Fig. 3.7). The discriminator *D2* in Fig. 3.7 learns to distinguish whether the composed image is realistic with respect to a set of *Real Images*. On the other hand, our study shows that real images are not good references for training geometry synthesizer. The reason is real images are realistic in both geometry and appearance spaces while the geometry can only achieve realism in geometry space. The difference in appearance space between the synthesized images and real images will mislead the training of geometry synthesizer. For optimal training of geometry synthesizer, the reference images should be realistic in the geometry space only and concurrently have similar appearance (e.g. colors and styles) with the initially composed images. Such reference images are difficult to create manually. In the SF-GAN, we elegantly use images from the appearance synthesizer (*Adapted Real* shown in Fig. 3.7) as the reference to train the geometry synthesizer, more details about the appearance synthesizer to be discussed in the following subsection.

3.4.2 Appearance Synthesizer

The appearance synthesizer is designed in a cycle structure as highlighted in orange-color lines and boxes on the right of Fig. 3.7. It aims to fuse the foreground object and background image to achieve synthesis realism in the appearance space. Image-to-image translation GANs also strive for realistic appearance but they usually lose

visual details while performing the appearance transfer. Within the proposed SF-GAN, guided filters are introduced which help to preserve visual details effectively while working towards synthesis realism within the appearance space.

Cycle Structure: The proposed SF-GAN adopts a cycle structure for mapping between two domains, namely, the composed image domain and the real image domain. Two generators *G1* and *G2* are designed to achieve image-to-image translation in two reverse directions, *G1* from *Composed Image* to *Final Synthesis* and *G2* from *Real Images* to *Adapted Real* as illustrated in Fig. 3.7. Two discriminators *D1* and *D2* are designed to discriminate real images and translated images.

In particular, *D1* will strive to distinguish the adapted composed images (i.e. the *Composed Image* after domain adaptation by *G1*) and *Real Images*, forcing *G1* to learn to map from the *Composed Image* to *Final Synthesis* images that are realistic in the appearance space G2 will learn to map from *Real Images* to *Adapted Real* images, the images that ideally are realistic in the geometry space only but have similar appearance as the *Composed Image*. As discussed in the previous subsection, the *Adapted Real* from *G2* will be used as references for training the geometry synthesizer as it will better focus on synthesizing images with realistic geometry (as the interfering appearance difference has been compressed in the *Adapted Real*).

Image appearance transfer usually comes with detail loss. We address this issue from two perspectives. The fist is by adaptive combination of cycle loss and identity loss. Specifically, we adopt a weighted combination strategy that assigns higher weight to the cycle-loss for interested image regions while higher weight to the identify-loss for non-interested regions. Take scene text image synthesis as an example. By assigning a larger cycle-loss weight and smaller identity-loss to text regions, it ensures a multi-mode mapping of the text style while keeping the background similar to the original image. The second is by introducing guided filters into the cycle structure for detail preserving, more details to be described in the next subsection.

Guided Filter: Guided filter was designed to perform edge-preserving image smoothing. It influences the filtering by using structures in a guidance image As appearance transfer in most image-to-image-translation GANs tends to lose image details, we introduce guided filters (F as shown in Fig. 3.7) into the SF-GAN for

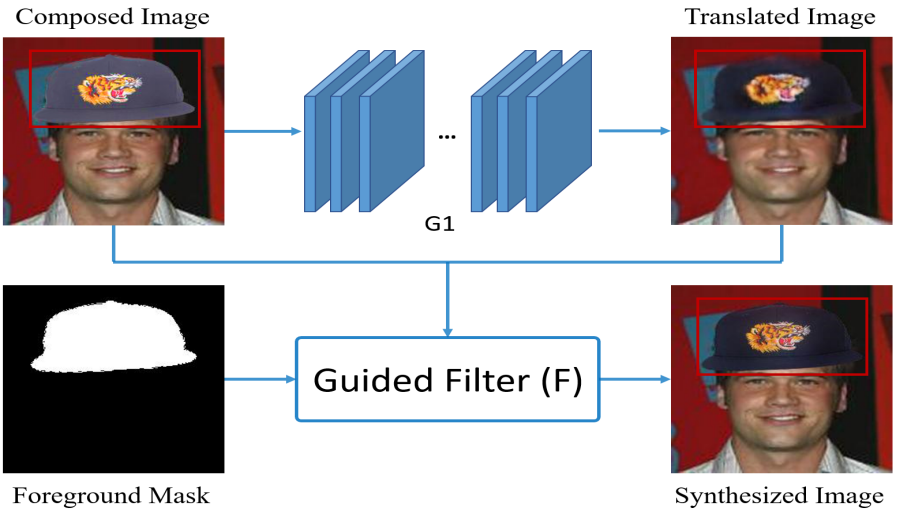


FIGURE 3.8: Detailed structure of the guided filter F: Given an image to be filtered (*Composed Image* in Fig. 3.7), a translated image with smoothed details (the output of *G1* in Fig. 3.7 where details are lost around background face and foreground hat areas) and the mask of foreground object hat (provided), F produces a new image with full details (*Synthesized Image*, the output of F at the bottom in Fig. 3.7). It can be seen that the guided filter preserves details of both background image (e.g. the face area) and foreground hat (e.g. the image areas highlighted by the red-color box).

detail preserving within the translated images. The target is to perform appearance transfer on the foreground object (within the *Composed Image*) only while keeping the background image with minimum changes.

We introduce guided filters into the proposed SF-GAN and formulate the detail-preserving appearance transfer as a joint up-sampling problem as illustrated in Fig. 3.8. In particular, the translated images from the output of *G1* (image details lost) is the input image I to be filtered and the initially *Composed Image* (image details unchanged) shown in Fig. 3.7 acts as the guidance image R to provide edge and texture details. The detail-preserving image T (corresponding to the *Synthesized Image* in Fig. 3.7) can thus be derived by minimizing the reconstruction error between I and T, subjects to a linear model:

$$T_i = a_k I_i + b_k, \forall i \in \omega_k \tag{3.3}$$

where i is the index of a pixel and ω_k is a local square window centered at pixel k.

To determine the coefficients of the linear model a_k and b_k, we seek a solution that minimizes the difference between T and the filter input R which can be derived by

minimizing the following cost function in the local window:

$$E(a_k, b_k) = \sum_{i \in \omega_k} ((a_k I_i + b_k - R_i)^2 + \epsilon a_k^2) \tag{3.4}$$

where ϵ is a regularization parameter that prevents a_k from being too large. It can be solved via linear regression:

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i \in \omega_k} I_i - \mu_k \overline{R}_k}{\overline{\sigma}_k + \epsilon}, \; b_k = \overline{R}_k - a_k \mu_k \tag{3.5}$$

where μ_k and σ_k^2 are the mean and variance of I in ω_k, $|\omega|$ is the number of pixels in ω_k, and $\overline{R}_k = \frac{1}{|\omega|}\sum_{i \in \omega_k}$ is the mean of R in ω_k.

By applying the linear model to all windows ω_k in the image and computing (a_k, b_k), the filter output can be derived by averaging all possible values of T_i:

$$T_i = \frac{1}{|\omega|} \sum_{k: i \in \mu_k} (a_k I_i + bk) = \overline{a}_i I_i + \overline{b}_i \tag{3.6}$$

where $\overline{a}_i = \frac{1}{|\omega|}\sum_{k \in \omega_k} a_k$ and $\overline{b}_i = \frac{1}{|\omega|}\sum_{k \in \omega_i} b_k$. We integrate the guide filter into the cycle structure network to implement an end-to-end trainable system.

3.4.3 Adversarial Training

The proposed SF-GAN is designed to achieve synthesis realism in both geometry and appearance spaces. The SF-GAN training therefore has two adversarial objectives, one is to learn the real geometry and the other is to learn the real appearance The geometry synthesizer and appearance synthesizer are actually two local GANs that are inter-connected and need coordination during the training. For presentation clarity, we denote the *Foreground Object* and *Background Image* in Fig. 3.7 as the x, the *Composed Image* as y and the *Real Image* as z which belongs to domains X, Y and Z, respectively.

For the geometry synthesizer, the *STN* can actually be viewed as a generator $G0$ which predicts transportation parameters for x. After the transformation of the *Foreground Object* and *Composition*, the *Composed Image* becomes the input of the discriminator $D2$ and the training reference $z^{'}$ comes from $G2(z)$ of the appearance synthesizer. For the geometry synthesizer, we adopt the Wasserstein

GAN [21] objective for training which can be denoted by:

$$\min_{G0} \max_{D2} E_{x\sim X}[D2(G0(x))] - E_{z'\sim Z'}[D2(z)] \tag{3.7}$$

where $Z^{'}$ denotes the domains for $z^{'}$. Since *G0* aims to minimize this objective against an adversary *D2* that tries to maximize it, the loss functions of *D2* and *G0* can be defined by:

$$L_{D2} = E_{x\sim X}[D2(G0(x)] - E_{z'\sim Z'}[D2(z^{'})], \quad L_{G0} = -E_{x\sim X}[D2(G0(x))] \tag{3.8}$$

The appearance synthesizer adopts a cycle structure that consists of two mappings $G1 : Y \rightarrow Z$ and $G2 : Z \rightarrow Y$. It has two adversarial discriminators *D1* and *D2*. D2 is shared between the geometry and appearance synthesizers, and it aims to distinguish y from *G2(z)* within the appearance synthesizer. The learning objectives thus consists of an adversarial loss for the mapping between domains and a cycle consistency loss for preventing the mode collapse. For the adversarial loss, the objective of the mapping $G1 : Y \rightarrow Z$ (and the same for the reverse mapping $G2 : Z \rightarrow Y$) can be defined by:

$$L_{D1} = E_{y\sim Y}[D1(G1(y)] - E_{z\sim Z}[D2(z)], \quad L_{G1} = -E_{y\sim Y}[D1(G1(y))] \tag{3.9}$$

As the adversarial losses cannot guarantee that the learned function maps an individual input y to a desired output z, we introduce cycle-consistency, aiming to ensure that the image translation cycle will bring x back to the original image, i.e. $y \rightarrow G1(y) \rightarrow G2(G1(y)) = y$. The cycle-consistency can be achieved by a cycle-consistency loss:

$$L_{G1_{cyc}} = E_{y\sim p(y)}[\|G2(G1(y)) - y\|], \quad L_{G2_{cyc}} = E_{z\sim p(z)}[\|G1(G2(z)) - z\|] \tag{3.10}$$

We also introduce the identity loss to ensure that the translated image preserves features of the original image:

$$L_{G1_{idt}} = E_{z\sim Z}[\|G1(z) - z\|], \quad L_{G2_{idt}} = E_{y\sim Y}[\|G2(y) - y\|] \tag{3.11}$$

For each training step, the model needs to update the geometry synthesizer and appearance synthesizer separately. In particular, L_{D2} and L_{G0} are optimized alternately while updating the geometry synthesizer. While updating the appearance synthesizer, all weights of the geometry synthesizer are freezed. In the mapping

$G1 : Y \rightarrow Z$, L_{D1} and $L_{G1} + \lambda_1 L_{G1_{cyc}} + \lambda_2 L_{G1_{idt}}$ are optimized alternately where λ_1 and λ_2 controls the relative importance of the cycle-consistency loss and the identity loss, respectively. In the mapping $G2 : Z \rightarrow Y$, L_{D2} and $L_{G2} + \lambda_1 L_{G2_{cyc}} + \lambda_2 L_{G2_{idt}}$ are optimized alternately.

It should be noted that the sequential updating is necessary for end-to-end training of the proposed SF-GAN. If discarding the geometry loss, we need update the geometry synthesizer according to the loss function of the appearance synthesizer. On the other hand, the appearance synthesizer will generate blurry foreground objects regardless of the geometry synthesizer and this is similar to GANs for direct image generation. As discussed before, the direct image generation cannot provide accurate annotation information and the directly generated images also have low quality and are not suitable for training deep network models.

3.4.4 Experiments

Datasets:

ICDAR2013 [159] is used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2013. It contains 848 images for network training and 1095 for testing.

ICDAR2015 [163] is used in the Robust Reading Competition under ICDAR 2015. It contains incidental scene text images that are captured without preparation before capturing. 2077 text image patches are cropped from this dataset, where a large amount of cropped scene texts suffer from perspective and curvature distortions.

IIIT5K [165] has 2000 training images and 3000 test images that are cropped from scene texts and born-digital images. Each word in this dataset has a 50-word lexicon and a 1000-word lexicon, where each lexicon consists of a ground-truth word and a set of randomly picked words.

SVT [166] is collected from the Google Street View images that were used for scene text detection research. 647 words images are cropped from 249 street view images and most cropped texts are almost horizontal.

TABLE 3.4: Scene text recognition accuracy over the datasets ICDAR2013, ICDAR2015, SVT, IIIT5K, SVTP and CUTE, where 1 million synthesized text images are used for all comparison methods as listed.

Methods	ICDAR2013	ICDAR2015	SVT	IIIT5K	SVTP	CUTE
Jaderberg [134]	58.1	35.5	67.0	57.2	**48.9**	35.3
Gupta [5]	62.2	38.2	48.8	59.1	38.9	36.3
Zhan [129]	**62.5**	37.7	63.5	59.5	46.7	36.9
ST-GAN [1]	57.2	35.3	63.8	57.3	43.2	34.1
SF-GAN(BS)	55.9	34.9	64.0	55.4	42.8	33.7
SF-GAN(GS)	57.3	35.6	66.5	57.7	43.9	36.1
SF-GAN(AS)	58.1	36.4	66.7	58.5	45.3	35.7
SF-GAN	61.8	**39.0**	**69.3**	**63.0**	48.6	**40.6**

SVTP [184] has 639 word images that are cropped from the SVT images. Most images in this dataset suffer from perspective distortion which are purposely selected for evaluation of scene text recognition under perspective views.

CUTE [185] has 288 word images most of which are curved. All words are cropped from the CUTE dataset which contains 80 scene text images that are originally collected for scene text detection research.

CelebA [186] is a face image dataset that consists of more than 200k celebrity images with 40 attribute annotations. This dataset is characterized by large quantities, large face pose variations, complicated background clutters, rich annotations, and it is widely used for face attribute prediction.

Scene Text Synthesis

Data Preparation: The SF-GAN needs a set of *Real Images* to act at references as illustrated in Fig. 3.7. We create the *Real Images* by cropping the text image patches from the training images of ICDAR2013 [159], ICDAR2015 [163] and SVT [166] by using the provided annotation boxes. While cropping the text image patches, we extend the annotation box (by an extra 1/4 of the width and height of the annotation boxes) to include certain local geometric structures

Besides the *Real Images*, SF-GAN also needs a set of *Background Images* as shown in Fig. 3.7. For scene text image synthesis, we collect the background images by smoothing out the text pixels of the cropped *Real Images*. Further, the *Foreground Object* (text for scene text synthesis) is computer-generated by using a 90k-lexicon.

TABLE 3.5: AMT user study to evaluate the realism of synthesized images. Percentages represent the how often the images in each category were classified as "real" by Turkers.

Methods	Text	Glass	Hat
Gupta et al. [5]	38.0	-	-
Zhan et al. [129]	41.5	-	-
ST-GAN [1]	31.6	41.7	42.6
Real	74.1	78.6	78.2
SF-GAN	57.7	62.0	67.3

The created *Background Images*, *Foreground Texts* and *Real Images* are fed to the network to train the SF-GAN.

For the training of scene text recognition model, texts need to be cropped out with tighter boxes (to exclude extra background). With the text maps as denoted by *Transformed Object* in Fig. 3.7, scene text patches can be cropped out accurately by detecting a minimal external rectangle.

Results Analysis: We use 1 million SF-GAN synthesized scene text images to train scene text recognition models and use the model recognition performance to evaluate the usefulness of the synthesized images. In addition, the SF-GAN is benchmarked with numbers of state-of-the-art synthesis techniques by randomly selecting 1 million synthesized scene text images from [134] and randomly cropping 1 million scene text images from [5] and [129]. Beyond that, we also synthesize 1 million scene text images with random text appearance by using ST-GAN [1]. There are many scene text recognition models [187–191], we design an attentional scene text recognizer with a 50-layer ResNet as the backbone network.

For ablation analysis, we evaluate SF-GAN(GS) which denotes the output of the geometry synthesizer (*Composed Image* as shown in Fig. 3.7) and SF-GAN(AS) which denotes the output of the appearance synthesizer with random geometric alignments. A baseline SF-GAN (BS) is also trained where texts are placed with random alignment and appearance. The three SF-GANs also synthesize 1 million images each for scene text recognition tests. The recognition tests are performed over four regular scene text datasets ICDAR2013 [159], ICDAR2015 [163], SVT [166], IIIT5K [165] and two irregular datasets SVTP [184] and CUTE [185] as described in **Datasets**. Besides the scene text recognition, we also perform user

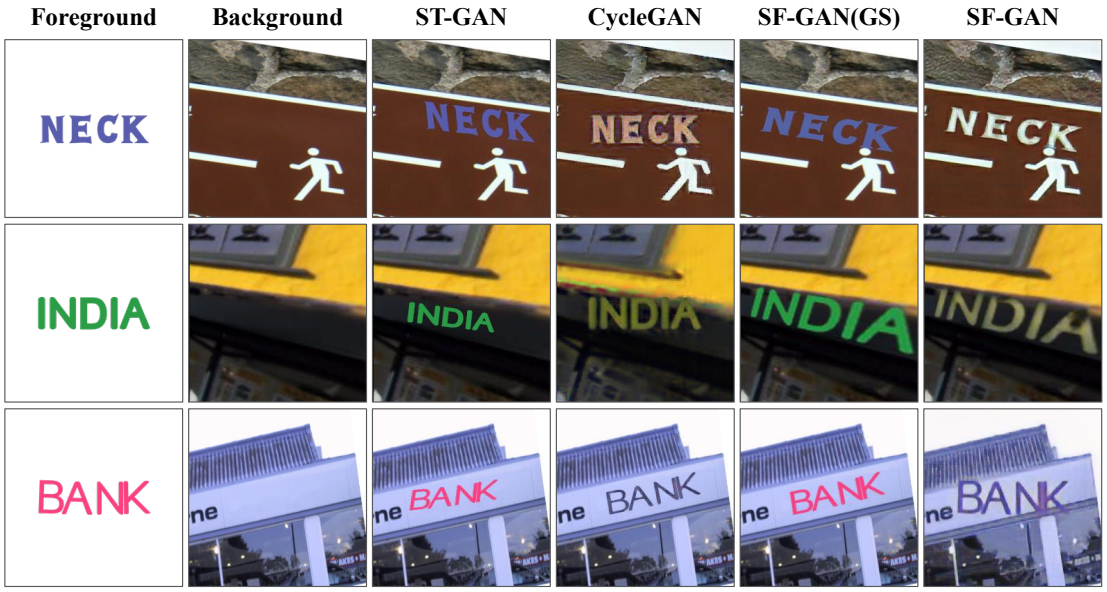


FIGURE 3.9: Illustration of scene text image synthesis by different GANs: Rows 1-2 are foreground texts and background images as labelled. Rows 3-4 show the images synthesized by ST-GAN and CycleGAN, respectively. Row 5 shows images synthesized by SF-GAN(GS), the output of the geometry synthesizer in SF-GAN (*Composed Image* in Fig. 3.7). The last row shows images synthesized by the proposed SF-GAN.

studies with Amazon Mechanical Turk (AMT) where users are recruited to tell whether SF-GAN synthesized images are real or synthesized.

Tables 3.4 and 3.5 show scene text recognition and AMT user study results. As Table 3.4 shows, SF-GAN achieves the highest recognition accuracy for most of the 6 datasets, demonstrating the superior usefulness of its synthesized images while used for training scene text recognition models. The ablation study shows that the proposed geometry synthesizer and appearance synthesizer both help to synthesize more realistic and useful image in recognition model training. In addition, they are complementary and their combination achieves a clear improvement in recognition accuracy beyond the baseline SF-GAN(BS). The AMT results in the second column of Table 3.5 also show that the SF-GAN synthesized scene text images are much more realistic than state-of-the-art synthesis techniques. Note the synthesized images by [134] are gray-scale and not included in the AMT user study.

Fig. 3.9 shows a few synthesis images by using the proposed SF-GAN and a few state-of-the-art GANs. As Fig. 3.9 shows, ST-GAN can achieve geometric alignment but the appearance is clearly unrealistic within the synthesized images.



FIGURE 3.10: Illustration of portrait-wearing by different GANs: Columns 1-2 show foreground hats and glasses and background face images, respectively. Columns 3-4 show images synthesized by by ST-GAN [1] and our proposed SF-GAN, respectively.

The CycleGAN can adapt the appearance of the foreground texts to certain degrees but it ignores real geometry. This leads to not only unrealistic geometry but also degraded appearance as the discriminator can easily distinguish generated images and real images according to the geometry difference. The SF-GAN (GS) gives the output of the geometry synthesizer, i.e. the *Composed Image* as shown in Fig. 3.7, which produces better alignment due to good references from the appearance synthesizer. In addition, it can synthesize curve texts due to the use of a thin plate spline transformation [183]. The fully implemented SF-GAN can further learn text appearance from real images and synthesize highly realistic scene text images. Besides, we can see that the proposed SF-GAN can learn from neighboring texts within the background images and adapt the appearance of the foreground texts accordingly.

Portrait Wearing

Data preparation: We use the dataset CelebA [186] and follow the provided training/test split for portrait wearing experiment. The training set is divided into two groups by using the annotation 'glass' and 'hat', respectively. For the glass case, one group of people with glasses serve as the real data for matching against in our adversarial settings and the other group without glasses serves as the background. For the foreground glasses, we crop 15 pairs of front-parallel glasses and reuse them

to randomly compose with the background images. According to our experiment, 15 pairs of glasses as the foreground objects are sufficient to train a robust model. The hat case has the similar setting, except that we use 30 cropped hats as the foreground objects.

Results Analysis: Fig 5. shows a few SF-GAN synthesized images and comparisons with ST-GAN synthesized images. As Fig. 3.10 shows, ST-GAN achieves realism in the geometry space by aligning the glasses and hats with the background face images. On the other hand, the synthesized images are unrealistic in the appearance space with clear artifacts in color, contrast and brightness. As a comparison, the SF-GAN synthesized images are much more realistic in both geometry and appearance spaces. In particular, the foreground glasses and hats within the SF-GAN synthesized images have harmonious brightness, contrast, and blending with the background face images. Additionally, the proposed SF-GAN also achieve better geometric alignment as compared with ST-GAN which focuses on geometric alignment only. We conjecture that the better geometric alignment is largely due to the reference from the appearance synthesizer. The AMT results as shown in the last two columns of Table 3.5 also show the superior synthesis performance of our proposed SF-GAN.

3.5 Illumination-Aware Image Composition

In this work, we design a novel Adversarial Image Composition Net (AIC-Net) [34] [6] that aims to achieve realistic image composition with adaptive shadow effects as illustrated in Fig. 3.11. AIC-Net adopts generative adversarial ideas to harmonize the foreground object within the background image and generates realistic shadow effects by learning from real shadows in natural images. It consists of local harmonization and global harmonization which are inter-connected and can be optimized jointly. We design a novel branched generation mechanism for high-fidelity local harmonization which disentangles the shadow generation and the style transfer of the foreground object for optimal accomplishment of two tasks simultaneously. An auxiliary illumination model is introduced to infer the illumination conditions of the background image such as lighting directions which help generate realistic shadows of the foreground object greatly. In addition, a differentiable spatial transformer

[6] The work has been published in ACCV 2020

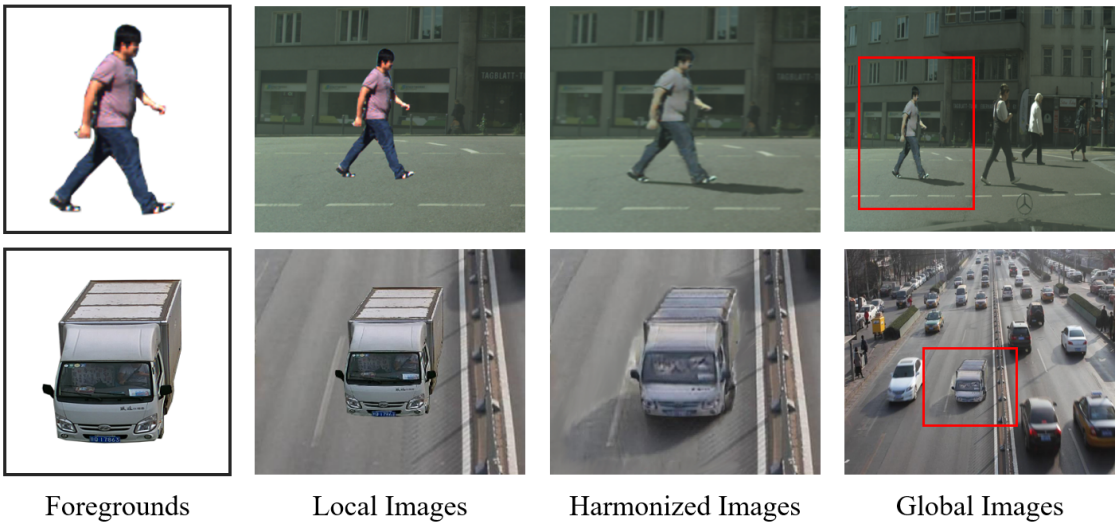


FIGURE 3.11: Given foreground objects and directly composed local images as shows in columns 1 and 2, the proposed AIC-Net disentangles the shadow generation and style transfer to produce realistic local harmonization as shown in column 3. With our spatial transformation module, the locally harmonized images in column 3 can be adapted and composed into the globally harmonized images as shown in column 4.

is introduced to bridge the local region and the global background for joint optimization of the local and global harmonization. Two discriminators are employed where a local discriminator guides to learn style transfer and shadow generation of the foreground object and a global discriminator guide to learn global image harmonization.

The contributions of this work are three folds. First, we design an innovative AIC-Net that is capable of generating high-fidelity shadow of the foreground object adaptively. Second, we design a novel branched generation mechanism that disentangles the style transfer and shadow generation of the foreground object and accomplishes the two tasks optimally and simultaneously. Third, we design spatial transformer module (STM) that achieves joint optimization of local harmonization and global harmonization effectively.

The proposed AIC-Net consists of local harmonization and global harmonization which are inter-connected and can be optimized jointly as illustrated in Fig. 3.12. Detailed network structures and training strategies will be presented in the following subsections.

3.5.1 Local Harmonization

The local harmonization is achieved by a local discriminator and a branched generation network which consists of a shadow branch and a texture branch as illustrated in Fig. 3.12. Given a background image (BG), a meaningful region within it is first selected for composition with a foreground object (FG). The selection of the meaningful region will be described in more details in *Experiment Setting*. We thus obtain a directly composed image (CI) x by embedding the foreground object into the selected region in the background image, as well as a foreground mask m_f by setting the foreground object pixels to 1 and the rest pixels in x to 0 as illustrated in Fig. 3.12

In local harmonization, the generator aims to transfer the style of the composed image and generate realistic object shadows concurrently. Although the two processes can be jointly handled through image-to-image translation, they have very different translation objectives and learn with distinct image-to-image mapping mechanisms. The concurrent learning of both degrades the translation severely as illustrated in 'W/o Branched Generation' in Fig. 3.13. We design a novel branched generation strategy to disentangle style transfer and shadow generation for optimal image harmonization. As shown in Fig. 3.12, the generator has two branches that share the same Encoder. With a background image and a composed image as the input, the decoder in shadow branch strives to generate shadows in the composed image while the decoder in the texture branch strives to transfer the style of the composed image to be compatible with the background image.

Shadow Branch: To avoid synthesizing shadows with false directions as shown in 'W/o Illumination Model' in Fig. 3.13, we need to infer the global illumination information from the background images. In this work, we adopt a learning-based method for illumination estimation. Specially, the ground truth of illumination map which is a panorama is represented by spherical harmonics (SH) coefficients [192], then a model with ResNet [193] as the backbone is adopted to regress the SH coefficients from the background images with limited field of view. With the predicted SH coefficients, an illumination map M_s can reconstructed as follows:

$$M(s) = \sum_{i=1}^{M} SH_i * y_i(s) \tag{3.12}$$

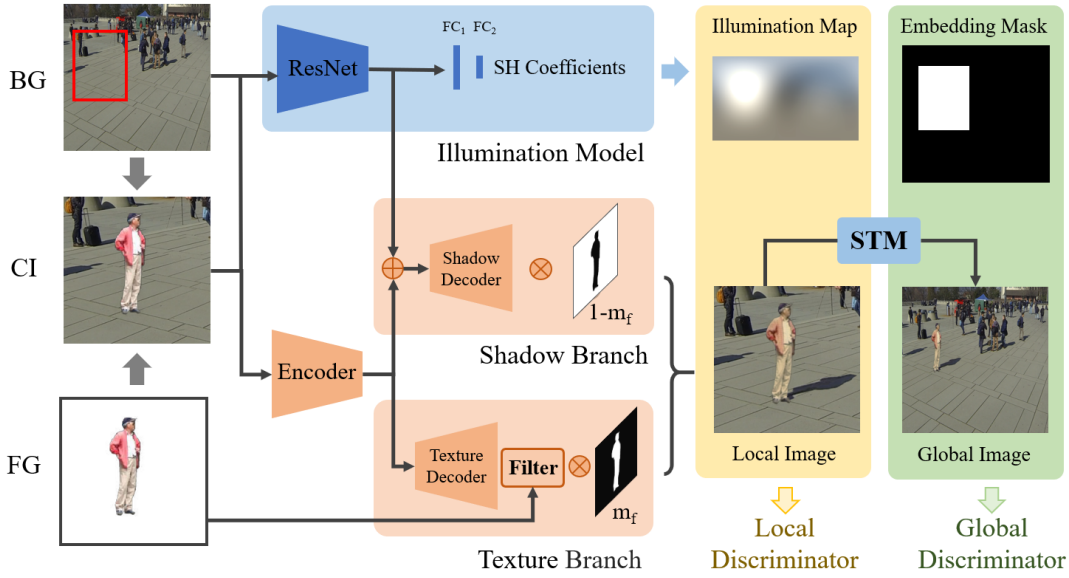


FIGURE 3.12: The structure of the proposed AIC-Net: With a background image *BG* (with specified composition regions) and foreground object *FG* to be composed, a directly composed images *CI* can be derived directly. A shadow branch and a texture branch will learn to generate shadow and transfer the style of *CI* which produces a harmonized *Local Image*. A spatial transformer module (STM) then learns to compose the *Local Image* with the background image to produce a *Global Image*. An *Illumination Model* will predict SH coefficients which can be used to reconstruct an *Illumination Map* ('FC' denote fully connected layer). The *Local Image* will be concatenated with the *Illumination Map* to be distinguished by the *Local Discriminator*. The *Global Image* will be concatenated with an *Embedding Mask* to be distinguished by the *Global Discriminator*.

where s is the spatial direction, M is the number of SH coefficients, y_i and SH_i are the i-th SH basis function and SH coefficients, respectively. We pre-train the illumination model on the Laval Indoor dataset [61] and [59] and fix the model weights while training AIC-Net.

The illumination model is able to infer the illumination from cues such as shadow existed in the background. Thus, the feature extracted by the illumination model contains rich information of shadow direction, size and intensity. In the shadow branch as shown in Fig. 3.12, we concatenate the feature extracted by the illumination model and the feature extracted by Encoder to integrate the shadow and illumination information into the shadow generation. The shadow decoder then strives to generate shadows with realistic shadow directions and styles in the local images. A binary mask (1-m_f) is applied to the output for integration with the texture branch. As the shadow branch may translate other regions beyond the

shadow areas undesirably, we feed real images into the shadow branch to keep the non-shadow areas be the same with the input. The shadow branch will thus concentrate on the translation of potential shadow areas for realistic shadow generation. To drive the effective learning of shadow branch, the reconstructed illumination map will be combined with the locally harmonized image to be distinguished by the local discriminator.

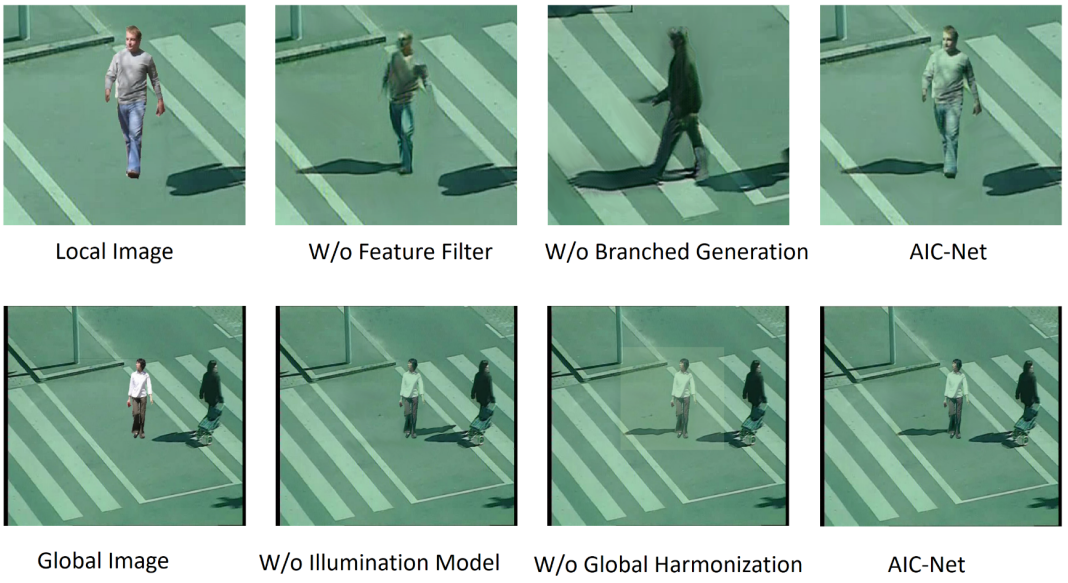


FIGURE 3.13: Illustration of the proposed guided feature filter, branched generation, illumination model and global harmonization: The introduction of the branched generation and guided feature filter helps to suppress various artifacts and generate realistic shadows and foreground object in the local image. The illumination model helps to generate shadow with high-fidelity shadow styles and directions. The global harmonization helps to suppress artifacts and style discrepancy between the local and global images.

Texture Branch: The texture branch aims to translate the texture of the foreground region to be harmonious with the background image. As a normal GAN generator tends to over-translate the foreground region which often degrades the foreground features as illustrated in 'W/o Feature Filter' in Fig. 3.13, we introduce a guided feature filter to preserve the 'content features' while performing style transfer. Different from the guided image filter [147] which is conducted in image level, guided feature filter is conducted in the feature space to achieve flexible style transfer. With the foreground object as the content guidance (C) and the feature decoded by the Texture Decoder as the style guidance (S), an output feature (T) with transferred style but preserved contents can be obtained through a linear

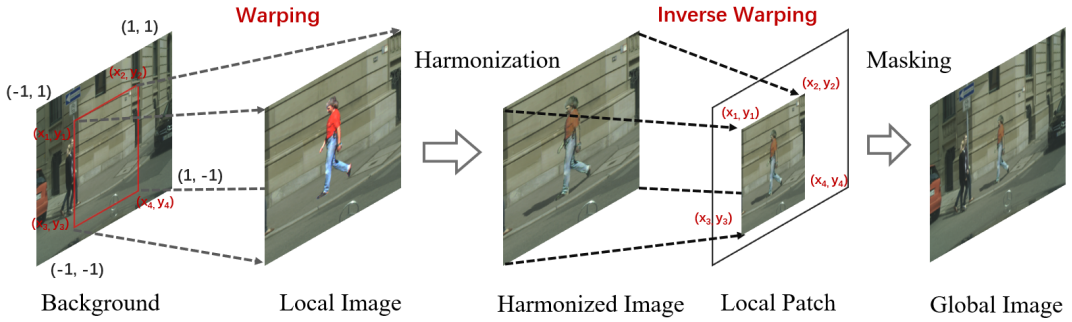


FIGURE 3.14: Global harmonization: With a selected region in the background image, a homography matrix H can be estimated to transform the selected region to the local image. After the local harmonization, the harmonized local image can be transformed back to the background image by an inverse warping H^{-1}. A masking operation is followed to obtain the global image.

model as follows:

$$T_i = a_k C_i + b_k, \forall i \in \omega_k \tag{3.13}$$

where i is a pixel index and ω_k is a local square window centered at pixel k, a_k and b_k are linear coefficients to be optimized. The minimization of the reconstruction error between C and T can be achieved by the following cost function:

$$E(a_k, b_k) = \sum_{i \in \omega_k} ((a_k S_i + b_k - C_i)^2 + \epsilon a_k^2) \tag{3.14}$$

where ϵ is a regularization parameter that prevents a_k from being too large. It can be solved via linear regression:

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i \in \omega_k} S_i - \mu_k \overline{C}_k}{\overline{\sigma}_k + \epsilon}, \quad b_k = \overline{R}_k - a_k \mu_k \tag{3.15}$$

where μ_k and σ_k^2 are the mean and variance of S in ω_k, $|\omega|$ is the number of pixels in ω_k, and $\overline{C}_k = \frac{1}{|\omega|}\sum_{i \in \omega_k} C_i$ is the mean of C in ω_k. Then we can obtain the transferred feature T. A following convolution layer will finally generate images with transferred style. A binary mask m_f is applied to the output for further integration with the shadow branch. By denoting the outputs of the shadow branch and the texture branch by x_s and x_t, the locally harmonized image x_h can be obtained as follows:

$$x_h = x_t * m_f + x_s * (1 - m_f) \tag{3.16}$$

3.5.2 Global Harmonization

The harmonized local region should be consistent with global image regarding the image style especially the boundary of the local region in the global image, otherwise the global image will tend to present clearly artifacts as shown in 'W/o Global Harmonization' of Fig. 3.13 which should be addressed as well for high-fidelity global harmonization. We adopt a global discriminator to ensure the consistency in global harmonization. With real background image as the training reference, the global discriminator learns to distinguish the harmonized global image (Fake Global in Fig. 3.12) and real global images (Real Global in Fig. 3.12), driving the generator for generating harmonious shadows and an attention network for eliminating the style discrepancy and boundary artifacts. A spatial transformer module (STM) is designed to bridge the local image and global image to achieve joint optimization.

Spatial Transformer Module: The spatial transformer is a differentiable module that applies a homography to the local image as illustrated in Fig. 3.14. We denote the coordinates of the four vertices of the background image by [(-1, 1), (1, 1), (1, -1), (-1, -1)], and the coordinate of the four vertices of the selected region by [(x1, y1), (x2, y2), (x3, y3), (x4, y4)] ($-1 < x_i < 1, -1 < y_i < 1, i = 1, 2, 3, 4$). With four paired vertices, the homograhy H can be computed which applied STM to warp the selected region to the local image. Once the local image is harmonized, it can be warped back to the local region of the background image by using the inverse homography H_{-1} as illustrated in Fig. 3.14. A warping mask with the selected region as 1 and other regions as 0 can be obtained simultaneously. A globally harmonized image can thus be obtained by masking the local patch into the background image. Since the STM and the masking operation are both differentiable, the local harmonization and the global harmonization can be bridged for end-to-end joint optimization.

3.5.3 Adversarial Training

With a GAN-based structure, the network is trained through adversarial learning between a generator and two discriminators. For clarity purpose, we denote the *BG*, *CI* in Fig. 3.12 as X and x, the generator and the two branches *Shadow*

Branch and *Texture Branch* as G, G_s and G_t, **Global Discriminator** and **Local Discriminator** as D_G and D_L. The foreground mask is denoted as m_f, the real global and local images as Y and y. The **Embedding Mask** in real global images and harmonized global images as m_Y and m_X.

For the local harmonization, the shadow and texture branches strive to generate shadow and translate foreground that are denoted by $G_s(X, x)$ and $G_t(X, x)$, respectively. The local image x_h can thus be denoted by:

$$x_h = G_s(X, x) * (1 - m_f) + G_t(X, x) * m_f \tag{3.17}$$

The local discriminator tries to distinguish the local image while the generator tries to mislead the local discriminator. As we adopt Wasserstein GAN [21] objective for training, the local adversarial loss of the generator and local discriminator can be defined by:

$$L_{D_L} = E[D_L(x_h)] - E[D_L(y)], \quad L_{G_L} = -E[D_L(x_h)] \tag{3.18}$$

To ensure the irrelevant region keep unchanged and preserve more details of the original image, an identity loss is defined as follows:

$$L_{S_{idt}} = E[G(Y, y) - y] \tag{3.19}$$

For the global harmonization, the global discriminator will distinguish the pair of global image X_h and embedding mask m_X which drives the generator to generate consistent texture and suppress artifacts around the local image boundary. The global adversarial loss of the generator and the global discriminator can thus be defined by:

$$L_{D_G} = E[D_G(X_h, m_X)] - E[D_G(Y, m_Y)], \quad L_{G_G} = -E[D_G(X_h, m_X)] \tag{3.20}$$

The overall objective function for the generator is:

$$L_G = L_{G_L} + \lambda_G L_{G_G} + \lambda_{G_{idt}} L_{S_{idt}} \tag{3.21}$$

where $\lambda_G, \lambda_{G_{idt}}$ denote weights of global adversarial loss and generator identity loss respectively. The overall objective function for the discriminators is:

$$L_D = L_{D_L} + \lambda_{D_G} L_{D_G} \tag{3.22}$$

where λ_{D_G} denotes the weight of global adversarial loss for the discriminator. L_D and L_G drive the learning of the whole model alternately.

3.5.4 Experiments

Dataset:

PRID 2011 [194] was created to evaluate person re-identification methods. The dataset consists of images extracted from multiple person trajectories recorded from two surveillance cameras.

WILDTRACK [195] is multi-camera detection and tracking dataset of pedestrians in the wild, where the cameras' fields of view has large part overlap.

Penn-Fudan [196] is an image database that was used for pedestrian detection evaluation as reported in [1]. The images are taken from scenes around campus and urban streets. The objects we are interested in these images are pedestrians. Each image will have at least one pedestrian.

UA-DETRAC [197] dataset consists of 100 challenging video sequences captured from over 140 thousand frames of real-world traffic scenes. More than 1.2 million labeled bounding boxes of vehicles with information of occlusion, weather, vehicle category and truncation are annotated.

Cityscapes [198] dataset focuses on semantic understanding of urban street scenes. It has street scenes from 50 cities and provides semantic segmentation of 30 classes. 3475 images are annotated with bounding boxes of persons.

ShapeNet [199] is a large repository of 3D object models which index more than 3 million models of which 220 thousand are classified into 3,135 categories.

Experiment Setting: In the pedestrian harmonization experiment, 2000 background images are collected from [194] and [195] and 500 pedestrians are cropped from [196] as the foreground objects. In the car harmonization experiment, we

collect the background images from [197] and [198] which contain images captured under different illumination conditions and viewpoints. Random horizontal flipping is applied to the background images to augment the training set. For the foreground images, 1000 3D car models from [199] are utilized to render 3000 car images in different views.

The local region to compose with the foreground object is derived by a simple mechanism. As the bounding boxes of the foreground objects are know, we first select a box region around an existed bounding box and the one with no overlap with other bounding boxes is selected as the local region.

The proposed method is implemented using the Tensorflow framework. The optimizer is Adadelta which employs adaptive learning rate. The model is trained in 100 thousand iterations with a batch size of 4. In addition, the network is trained on a workstation with an Intel Core i7-7700K CPU and a NVIDIA GeForce GTX 1080 Ti graphics card with 12GB memory. All input images are resized to 256×256 in training.

TABLE 3.6: Comparison of harmonization methods with different metrics: 'Car' and 'Ped' denote the results of car and pedestrian synthesis respectively. 'Local' and 'Global' denote harmonized local and global images. 'MS' metric denotes the manipulation scores. For AMT user study, the number in each cell is the percentage of the harmonized images that are deemed as the most realistic by Turkers.

Metrics	**CycleGAN**		**DeepIH**		**DeepPH**		**AdaIN**		**AIC-Net**	
	Car	Ped	Car	Ped	Car	Ped	Car	Ped	Car	Ped
AMT(Local)	8%	4%	12%	11%	20%	23%	18%	17%	**42%**	**45%**
AMT(Global)	3%	1%	16%	15%	25%	24%	21%	22%	**41%**	**42%**
FID(Local)	115.1	131.6	136.2	151.4	99.1	103.3	88.7	97.2	**84.4**	**97.1**
FID(Global)	102.4	126.1	121.8	139.3	92.4	101.9	84.6	92.8	**72.2**	**89.6**
MS(Local)	0.69	0.70	0.75	0.74	0.72	0.73	0.69	0.71	**0.63**	**0.64**
MS(Global)	0.71	0.70	0.73	0.76	0.73	0.74	0.70	0.74	**0.62**	**0.62**

Experiment Analysis

Quantitative Analysis: The evaluation is performed through Frechet Inception Distance (FID) [200], Manipulation Score (MS) [123] and Amazon Mechanical Turk (AMT) study. For the evaluation with FID, the harmonized images are fed to the Inception network [201] trained on ImageNet [202], and the features from the layer before the last fully connected layer are used to calculate the Fréchet Inception

Distance. The Manipulation Score [123] is generated by a state-of-the-art manipulation detection algorithm [203] where a higher score indicates a higher possibility that the image is manipulated. For the AMT study which involves 10 users, we present a set of images (5 images by 5 methods) each time to the user to select the most realistic one. With 20 sets of presented images, an AMT score is derived by averaging the percentage of most realistic images as judged by the 10 users. 'Local' and 'Global' denote the results of locally and globally harmonized images respectively.

Table 3.6 compares the quality and realism of images processed by different harmonization methods. For the local harmonization, AIC-Net outperforms all compared methods with all metrics clearly as it harmonizes the foreground region and generates relevant shadow which improves the image realism significantly. The using of guided filter helps to preserve texture of high quality, thus obtaining the best FID scores. The spatial transformer module achieves the joint optimization of local and global harmonization, leading to best Manipulation score. CycleGAN can also synthesize shadows but the foreground region tends to be translated undesirably as it does not disentangle the shadow generation and style transfer processed. As DeepIH requires paired training images with semantic annotations, it performs terribly bad when paired training images are not available. As style transfer methods, DeepPH and AdaIN harmonize the foreground region successfully but all of them ignore the shadow generation completely. For the global harmonization, AIC-Net outperforms other methods by large margins as its joint optimization of local and global harmonization ensures consistent shadow and smooth local boundary. The lack of global harmonization leads to clear artifacts in the global image harmonized by CycleGAN. For DeepIH, DeepPH and AdaIN, there are not many artifacts but the lack of shadow and guided filter degrades the image realism and quality greatly. Normally, the global harmonized images present better score than local harmonized images as only small part of global image needs to be harmonized.

Qualitative Analysis: We compare the images harmonized by different methods in Fig. 3.15, where AIC-Net (L) and AIC-Net (G) denote harmonization of local images and global images by our proposed AIC-Net. As Fig. 3.15 shows, DeepIH [46] can only achieve a slight translation and cannot generate any potential shadow, thus resulting in unrealistic harmonized images. The branched generation in AIC-Net disentangles the style transfer and shadow generation, leading to much more

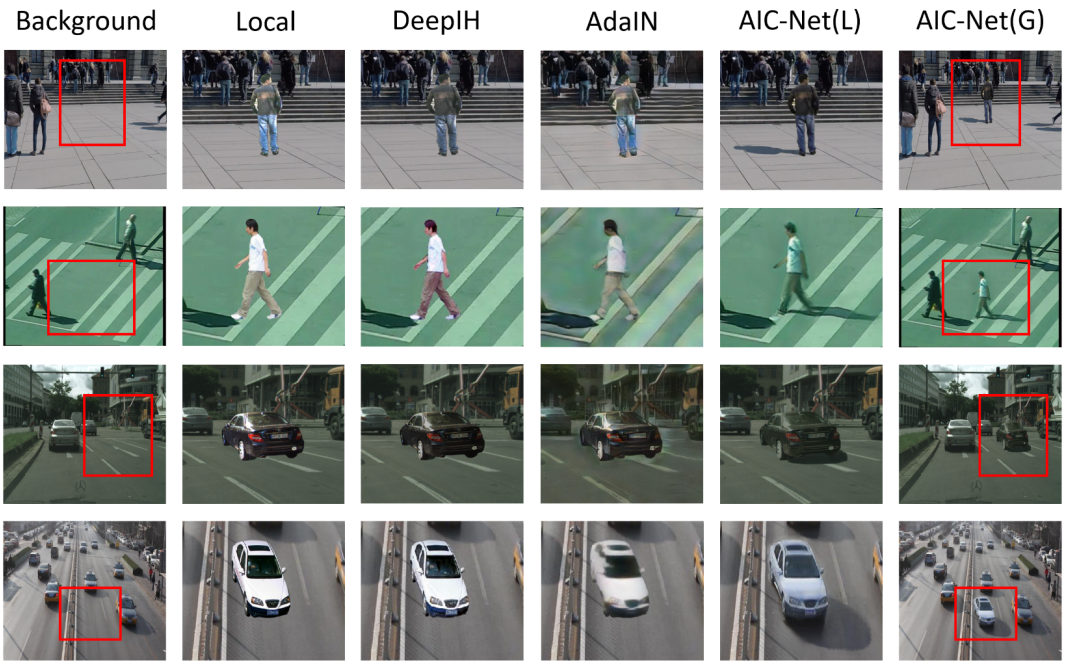


FIGURE 3.15: Comparing AIC-Net with other harmonization methods: Images in col. 1 are background images with selected regions. Col. 2 shows directly composed local images. AIC-Net (L) and AIC-Net (G) show images after local and global harmonization.

realistic image harmonization. The global discriminator in AIC-Net works better by driving the generator to remove the boundary artifacts. AdaIN [204] transfer the style of the foreground region to be harmonious with the background, but they do not handle shadow generation which leads to unrealistic composition. Besides, AdaIN is unable to preserve the texture quality of harmonized image due to the missing of guided filter.

To validate that the proposed model is able to generate various shadow adaptively, we tested the trained model in different environments as shown in Fig. 3.16. The first two rows show the backgrounds and local images respectively, the last row shows locally harmonized images and globally harmonized images produced by our model. As shown in the first three columns of Fig. 3.16, the proposed model generates shadows conditioned on the existed shadow cases in the backgrounds. For background images without clearly shadow case as shown in the last two columns of Fig. 3.16, the proposed model is still able to synthesize realistic shadows conditioned on the feature extracted by the illumination model.

Ablation Study: We conducted an ablation study to evaluate the effectiveness of

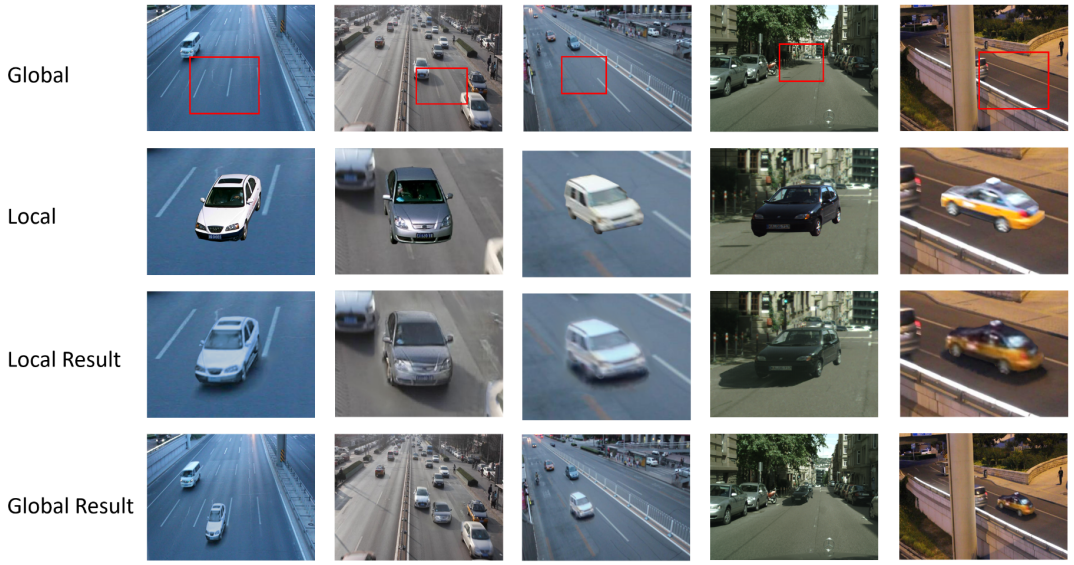


FIGURE 3.16: Shadow generation under different conditions by AIC-Net: The first row shows the background image with the local region for foreground embedding. The second row show the direct composition. Rows 3-4 show the local harmonization and global harmonization by AIC-Net, respectively. AIC-Net can generate various shadows adaptively conditioned on illumination, viewpoint and weather of the background image, etc.

TABLE 3.7: We conduct ablation study with three metrics (FID, AMT, MS) to evaluate the quality of harmonized local and global images. AIC-Net (WF), AIC-Net (WB), AIC-Net (WS) and AIC-Net(WI) denote AIC-Net without the guided feature filter, without branched generation mechanism, without spatial transformer module and without illumination model respectively. For AMT user study, the percentage numbers represent how often the harmonized images in each category are classified as *most realistic* by Turkers. We average the FID, AMT and MS scores of car and pedestrian images as the final scores.

Methods	**FID ↓**		**AMT ↑**		**MS ↓**	
	Local	Global	Local	Global	Local	Global
AIC-Net(WF)	115.5	106.1	14%	19%	0.72	0.70
AIC-Net(WB)	109.2	93.7	19%	20%	0.70	0.69
AIC-Net(WS)	98.5	92.1	21%	16%	0.73	0.71
AIC-Net(WI)	94.2	91.8	17%	18%	0.67	0.65
AIC-Net	**90.8**	**81.8**	**29%**	**27%**	**0.63**	**0.62**

different technical designs in AIC-Net. As shown in Table 3.7, four variants of AIC-Net are designed including: 1) AIC-Net (WF) denotes AIC-Net without the guided feature filter; 2) AIC-Net (WB) denotes AIC-Net without using the branched generation mechanism; 3) AIC-Net (WS) denotes AIC-Net without incorporating the spatial transformer module; 4) AIC-Net(WI) denotes AIC-Net without illumination model. The evaluations are performed with three metrics including FID, MS

and AMT study that involves 10 users. For AMT user study, a set of images (5 generated by AIC-Net and its four variants) are presented to each of 10 users who will choose the most realistic image. We presented 25 sets of harmonized images and the AMT score is derived by averaging the percentages of the most realistic images as judged by the 10 users. The scores of car and pedestrian experiments are averaged for each metric to obtain the final scores.

As Table 3.7 shows, AIC-Net (WF) obtains clearly lower scores with all metrics as compared with the standard AIC-Net which shows the importance of guided feature filter in generating realistic image harmonization. AIC-Net (WB) also obtains lower scores than AIC-Net as the branched generation in AIC-Net disentangles the shadow generation and style transfer to produce more realistic local harmonization as shown in Fig. 3.13. For AIC-Net (WS), the scores for the global images are much lower than the other models, demonstrating the importance of the spatial transformer module which bridges the local and global harmonization to produce globally realistic harmonization. On the other hand, AIC-Net (WS) achieves similar scores which can be expected as both models are almost the same in local harmonization. AIC-Net (WI) present clearly drop of AMT score compared with standard AIC-Net, indicating the significance of illumination model for the generation of realistic shadow.

3.6 Conclusion

This chapter presents a scene text image synthesis technique that aims to train accurate and robust scene text detection and recognition models. The proposed technique achieves verisimilar scene text image synthesis by combining three novel designs including semantic coherence, visual attention, and adaptive text appearance. Experiments over 5 public benchmarking datasets show that the proposed image synthesis technique helps to achieve state-of-the-art scene text detection and recognition performance.

In addition, we present SF-GAN that synthesizes realistic images given foreground objects and background images. SF-GAN can achieve synthesis realism in both geometry and appearance spaces concurrently. The first scene text image synthesis study shows that SF-GAN can synthesize useful images to train better recognition

models. The second portrait-wearing study shows the SF-GAN is widely applicable and can be easily extended to other tasks.

Finally, we present AIC-Net, an end-to-end trainable network that harmonizes foreground objects and background images with adaptive and realistic shadow effects. A branched generation mechanism is designed to disentangle style transfer and shadow generation to produce high-fidelity local harmonization. An auxiliary illumination model is incorporated in shadow branch to achieve adaptive shadow generation. A spatial transformer module is introduced to bridge the local and global harmonization for joint optimization. Quantitative and qualitative experiments demonstrate the superior harmonization performance and wide adaptability of the proposed method.

Chapter 4

Translation-based Image Synthesis

4.1 Introduction

Adversarial image synthesis has achieved remarkable improvements with the advance of convolutional neural networks (CNNs) in recent years. Most generative approaches model image synthesis pixel-wise by learning a mapping function from latent vectors to the subspace of real images, which can be referred as direct image generation. Generative Adversarial Networks (GANs) [17] has demonstrated its superior performance in realistic image generation through a min-max game between a generator network that produces images from codes and a discriminator network (D) that distinguishes real images from the generated ones until reaching the equilibrium. However, the direct generation methods mainly work well on restricted scenes since the finite network capacity makes it hard to fit the high-dimension space of real images. In this chapter, we present to utilize Spatial Transformer Networks (STNs) [205] to achieve high-fidelity image synthesis via geometric transformation. Specifically, we explored the translation of image geometry for the tasks of scene text rectification and domain adaptation.

Accurate and robust recognition of scene texts by machines has been a research challenge for years, largely due to huge amounts of variations in text appearance, the complicated image background, imaging artifacts, etc. The advance of deep learning research and its successes in many computer vision tasks have pushed the

boundary of scene text recognition greatly in recent years [134, 162, 174, 175, 188, 206, 207]. On the other hand, the learning-based approach is still facing various problems while dealing with a large amount of scene texts that suffer from arbitrary perspective distortions and text line curvature. We thus designed an innovative text recognition network ESIR [35] [1] that rectifies geometric distortions of scene texts iteratively by translating scene texts from perspective views to fronto-parallel views, more details to be described in Section 4.3.

A large amount of labelled or annotated images is critical for training robust and accurate deep neural network (DNN) models, while collecting and annotating large datasets are often extremely expensive. The lack of scalability and transferability of CNN models makes data collection and annotation even more challenging while dealing with images collected under different conditions from different domains. Unsupervised Domain Adaptation (DA), which transfers images and features from a source domain to a target domain, has achieved very impressive performance especially with the recent advances of Generative Adversarial Networks (GANs) [17]. Different DA techniques have been developed and applied to various computer vision problems successfully such as style transfer, image synthesis, etc. State-of-the-art DA still faces various problems. In particular, most existing systems focus on learning feature shift in appearance space only whereas the feature shift in geometry space is largely neglected. On the other hand, images from different domains often differ in both appearance and geometry spaces. Take various texts in scenes as an example. They could suffer from motion blurs in appearance space and perspective distortion in geometry space concurrently as shown in the target domain images in Fig. 4.6, and both are essential features for learning robust and accurate scene text detectors and recognizers. As a result, existing techniques often suffer from a clear performance drop when source and target domains have clear geometry discrepancy as observed for images and videos from different domains. We designed a geometry-aware domain adaptation network [36] [2] for scene text detection and recognition, more details to be described in Section 4.4.

[1] The work has been published in CVPR 2019
[2] The work has been published in ICCV 2019

4.2 Related Works

4.2.1 Distorted Scene Texts Recognition

The state-of-the-art combining RNNs and attention has achieved great success while dealing with horizontal or slightly distorted texts in scenes. On the other hand, most existing methods still face various problems while dealing with many scene texts that suffer from either perspective distortions or text line curvatures or both. Prior works dealing with perspective distortions and text line curvatures are limited but this problem has attracted increasing attention in recent years. The very early works [208–210] correct perspective distortions in document texts as captured by digital cameras for better recognition. [184] works with scene texts by using bag of key-points that are tolerant to perspective distortions. These early systems achieve limited successes as they use hand-crafted features and require character-level information. The recent works [162, 190] also take an image rectification approach but explore spatial transformer networks for scene text distortion correction. Similarly, [211, 212] integrate the rectification and recognition into the same network. These recent systems exploit deep convolutional networks for rectification and RNNs for recognition, which require little manually crafted features or extra annotations and have shown very promising recognition performance.

Our proposed technique adopts a rectification approach for robust and accurate recognition of scene texts with perspective and curvature distortions. Different from existing rectification based works [162, 190, 211, 212], it corrects distortions in an iterative manner which helps to improve the rectification and recognition greatly. In addition, we propose a novel line-fitting transformation that is robust and flexible in scene text distortion estimation and correction.

4.2.2 Domain Adaptation

Domain adaptation is an emerging research topic that aims to address domain shift and dataset bias [213, 214]. Existing techniques can be broadly classified into two categories. The first category focuses on minimizing discrepancies between a source domain and a target domain in the feature space. For example, [215] explored Maximum Mean Discrepancies (MMD) and Joint MMD distance across

domains over fully connected layers. [216] studied feature adaptation by minimizing the correlation distance and then extended it to deep architectures [217]. [218] modelled domain-specific features to encourage networks to learn domain-invariant features. [219, 220] improved feature adaptation by designing various adversarial objectives.

The second category adopts Generative Adversarial Nets (GANs) [17] to perform pixel-level adaptation via continuous adversarial learning between generators and discriminators which has achieved great success in image generation [18, 19, 110], image composition [1] and image-to-image translation [22, 25, 113]. Different approaches have been investigated to address pixel-level image transfer by enforcing consistency in the embedding space. [221] translates a rendering image to a real image by using conditional GANs. [222] studies an unsupervised approach to learn pixel-level transfer across domains. [182] proposes an unsupervised image-to-image translation framework using a shared-latent space. [223] introduces an inference model that jointly learns a generation network and an inference network. More recently, CycleGAN [113] and its variants [224, 225] achieve very impressive image translation by using cycle-consistency loss. [181] proposes a cycle-consistent adversarial model that adapts at both pixel and feature levels.

4.3 Iterative Geometry Transformation

We design an end-to-end trainable scene text recognition network ESIR [35] [3] that performs instance-level image translation by rectifying scene texts from perspective views to fronto-parallel views. ESIR employs an innovative rectification network that corrects perspective and curvature distortions of scene texts iteratively as illustrated in Fig. 4.1. The finally rectified scene text image is fed to a recognition network for recognition. The training of the iterative rectification network is driven by better scene text recognition as back propagated from the recognition network, requiring no other annotations beyond the scene texts as used in most scene text recognition systems.

[3]The work has been published in CVPR 2019

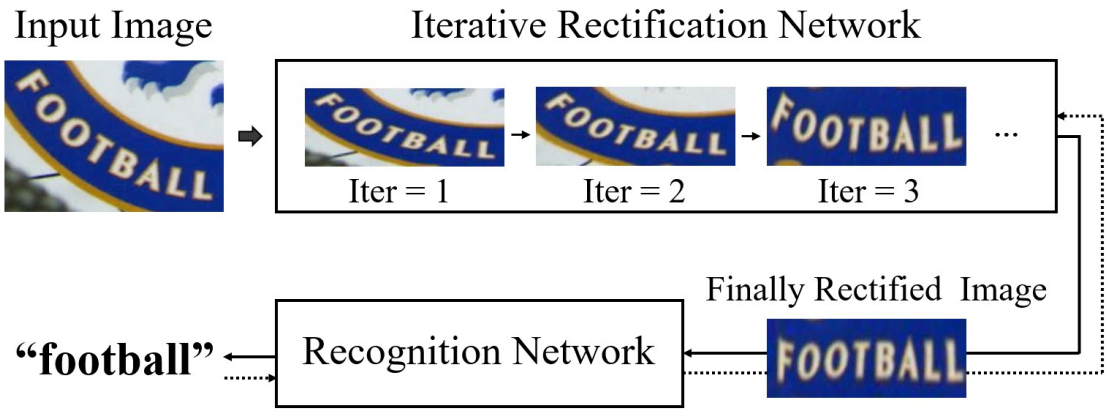


FIGURE 4.1: The proposed system consists of an Iterative Rectification Network that corrects scene text distortions iteratively and a Recognition Network that recognizes the finally rectified scene texts (the images within the Iterative Rectification Network illustrate the iterative scene text rectification process). It is end-to-end trainable as driven by better scene text recognition performance, where the dotted lines show the back propagation of gradients.

The proposed ESIR addresses two typical constraints in the scene text recognition problem. The first is *robust* distortion rectification for optimal scene text recognition. To address this challenge, we design a novel line-fitting transformation that is powerful and capable of modeling and correcting various scene text distortions reliably. The line-fitting transformation models the middle line of scene texts by using a polynomial which is able to estimate the pose of either straight or curved text lines flexibly as illustrated in Fig. 4.2. In addition, it employs line segments which are capable of estimating the orientation and the boundary of text lines in vertical direction reliably. The proposed rectification network is thus capable of correcting not only perspective distortions in straight text lines as in spatial transfer networks [205] and bag-of-keypoints recognizer [184], but also various curvatures in curved text lines in scenes.

The second is *accurate* rectification of perspective and curvature distortions of texts in scenes. To address this challenge, we develop an iterative rectification pipeline that employs multiple feed-forward rectification modules to estimate and correct scene text distortions iteratively. As illustrated in Fig. 4.3, each iteration takes the image rectified in the last iteration for further distortion estimation as driven by higher scene text recognition accuracy. The iterative rectification is thus capable of producing more accurate distortion correction compared with the state-of-the-art that just performs a single distortion estimation and correction [162, 190]. In addition, the iteratively rectified images lead to superior scene text recognition

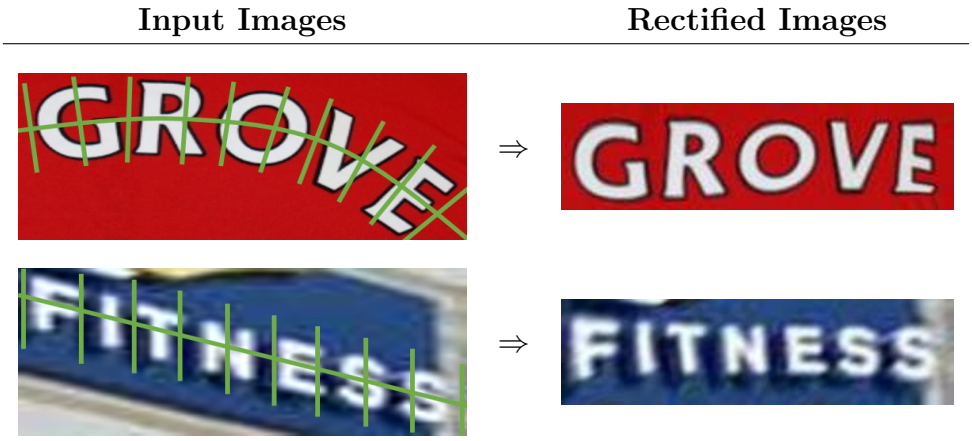


FIGURE 4.2: A novel line-fitting transformation is designed which employs a polynomial to model the middle line of scene texts in horizontal direction, and a set of line segments to estimate the orientation and the boundary of text lines in vertical direction. It helps to model and correct various scene text distortions accurately.

accuracy especially for datasets that contain a large amount of curved and/or perspectively distorted texts, to be described in experiments.

The contributions of this work are threefold. First, it proposes a novel line-fitting transformation that is flexible and robust for scene text distortion modeling and correction. Second, it designs an iterative rectification framework that clearly improves the scene text rectification and recognition performance with no extra annotations. Third, it develops an end-to-end trainable system that is robust to parameter initialization and achieves superior scene text recognition performance across a number of public datasets.

The following sections present the proposed techniques including iterative scene text rectification network, sequence recognition network and detailed description of network training.

4.3.1 Rectification Network

The proposed iterative rectification network employs a novel line-fitting transformation and an iterative rectification pipeline for optimal estimation and correction of perspective and curvature distortions of texts in scenes.

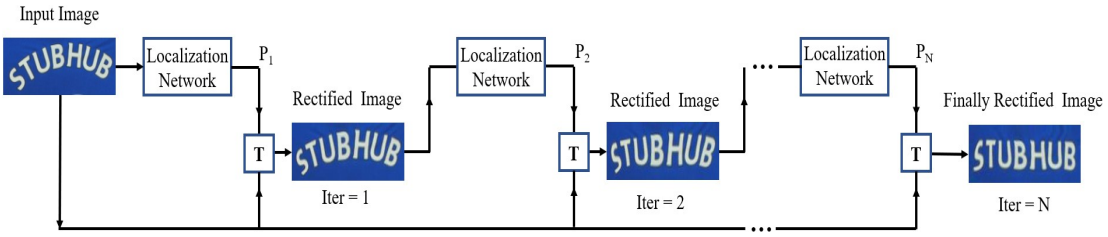


FIGURE 4.3: The iterative rectification process: T denotes a thin plate spline transformation, P_1, P_2, ... denote transformation parameters that are predicted by the localization network, Iter denotes the number of rectification iterations and N is the predefined iteration number.

Line-Fitting Transformation: A novel line-fitting transformation is designed to model the pose of scene texts and correct perspective and curvature distortions for better scene text recognition. As illustrated in Fig. 4.2, the fitting lines consist of a polynomial that is employed to model the middle line of text lines in horizontal direction, and L line segments that are employed to estimate the orientation and the boundary of text lines in vertical direction. Since most text lines in scenes are either along a straight or smoothly curved line, a polynomial of a certain order is sufficient for the estimation of text line poses in the horizontal direction. In our implemented system, a polynomial of order 4 and a set of 10 line segments are employed for scene text pose estimation.

By setting the image center as the origin and normalizing the x-y coordinate of each pixel within scene text images, the middle line of text lines in scenes can be modeled by a polynomial of order K as follows:

$$y = a_K * x^K + a_{K-1} * x^{K-1} + \cdots + a_1 * x + a_0$$

The L line segments can be modeled by:

$$y = b_{1,l} * x + b_{0,l} \mid r_l, \quad l = 1, 2, \cdots, L$$

where r_l denotes the length of line segments on the two sides of the middle line of text lines in scenes which can be approximated as the same. We therefore have $3L$ parameters for estimating the L line segments. By including the middle line polynomial, the number of parameters to be estimated becomes $3L + K + 1$.

The proposed rectification network iteratively regresses to estimate the fitting-line parameters by employing a localization network together with image convolutions

as illustrated in Fig. 4.3. It should be noted that the training of the localization network does not require any extra annotation of fitting lines but is completely driven by the gradients that are back propagated from the recognition network. The underlying principle is that higher recognition performance is usually achieved when scene text distortions are better estimated and corrected.

Once the fitting line parameters are estimated, the coordinates of 2 endpoints of each of the L line segments $\{t_j|j=1,\cdots,2L\}$ can be determined. Scene text distortions can then be corrected by a thin plate spline transformation [183] that can be determined based on the estimated line segment endpoints and another $2L$ base points $\left\{t'_j|j=1,\cdots,2L\right\}$ which define the appearance of scene texts within the rectified scene text image:

$$t'_j = \begin{cases} [-0.5+\frac{j-1}{L-1}, 0.5]^T, & 1 \le j \le L \\ [-0.5+\frac{l-(L+1)}{L-1}, -0.5]^T, & L+1 \le j \le 2L \end{cases}$$

With $P=[t_1, t_2, \cdots t_{2L}]^T$ and $P'=[t'_1, t'_2, \cdots t'_{2L}]^T$, the transformation parameters can be determined by:

$$C = \begin{bmatrix} S & 1_{2L} & P \\ 1^T_{2L} & 0 & 0 \\ P^T & 0 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} P' \\ 0 \\ 0 \end{bmatrix}$$

where S=$[U(t\text{-}t_1),\ U(t\text{-}t_2),\ \cdots,\ U(t\text{-}t_{2L})]^T$ and $U(r)=r^2 log r^2$. For every pixel t' within the rectified scene text image, the corresponding pixel t within the distorted scene text image can thus be determined by: $t=C\cdot t'$. With the estimated pixels $\left\{t'_j|j=1,\cdots,2L\right\}$, a grid $G=\{t_j|j=1,\cdots,2L\}$ can be generated within the distorted scene text image for rectification. A sampler is implemented to produce the rectified scene text image by using the determined grid, where the value of the pixel t' is bilinearly interpolated from the pixels near t within the distorted scene text image. The sampler can back propagate the image gradients as it is fully differentiable.

Iterative Rectification: We develop an iterative rectification pipeline for optimal scene text rectification and recognition. At the first iteration, the rectification network takes the original scene text image as input and rectifies it to certain degrees by using the estimated transformation parameters as described in the last subsection. After that, the rectified scene text image is fed to the same rectification network for further parameter estimation and image rectification. This process repeats until a predefined number of rectification iterations is reached. The finally rectified image is then fed to the sequence recognition network for scene text recognition.

The iterative rectification improves the scene text image rectification performance greatly as to be described in experiments. On the other hand, it often encounters a critical 'boundary effect' problem if the iterative rectification is performed directly without control. In particular, each rectification iteration discards image pixels lying outside of the image sampling region, which accumulates and could lead to the discarding of certain text pixels during the iterative image rectification process. In addition, direct image rectification iteratively often degrades the image clarity severely because of the multiple rounds of bilinear interpolations.

We deal with the 'boundary effects' and losing of image clarity by designing a novel network structure as illustrated in Fig. 4.3. In particular, the rectification network consists of localization networks for estimating rectification parameters, and a thin plate spline transformation T that employs the estimated rectification parameters to generate rectified scene text images. During the iterative image rectification process, the intermediately rectified scene text image is used for parameter estimation only, and the original instead of intermediately rectified scene text image is fed to the transformation module T for rectification consistently. With this new design, 'boundary effect' accumulation can be avoided and image clarity will not be degraded, both help to improve the scene text recognition performance greatly as to be presented in experiments.

4.3.2 Recognition Network

The recognition network employs a sequence-to-sequence model with an attention mechanism. It consists of an encoder and a decoder. In the encoder, the input is a rectified scene text image that is re-sized to 32×100 pixels. A 53-layer residual

network [193] is used to extract features, where each residual unit consists of a 1×1 convolution and a 3×3 convolution operations. A 2×2 stride convolution is implemented to down-sample feature maps in the first two residual blocks. The convolution stride is then changed to 2×1 in all following residual blocks and this helps to reserve more information along the horizontal direction and is also very useful for distinguishing neighboring characters. The residual network is followed by two layers of Bidirectional long short-term memory (BiLSTM) each of which has 256 hidden units. The decoder adopts the LuongAttention mechanism [226] which consists of 2-layer attentional LSTMs with 256 hidden units and 256 attention units. During the inference stage, beam search is employed to decode components that maintain k candidates with the highest accumulative scores.

4.3.3 Network Training

The training of image rectification networks is never a simple task. The major issue is that image rectification networks are sensitive to parameter initialization as frequently encountered and shared in prior studies [162, 190]. In particular, random parameter initialization often leads to network convergence problems because it is liable to produce highly distorted scene text images that ruin the training of the recognition network and further the rectification network (note that the training of the rectification network is driven by the scene text recognition performance).

We address the network initialization problem by avoiding direction prediction of P as defined in **Line-Fitting Transformation**. Instead, we define an auxiliary P_0 that equals P' at the beginning and make $P = P_0 + \triangle P$, where $\triangle P$ is predicted by the rectification network iteratively. By assigning a small value to $\triangle P$, the initial P will have a similar value as P'. This approach avoids generating highly distorted scene text image at the beginning stage and improves the network convergence greatly. Additionally, the gradual learning of P via iterative estimation of $\triangle P$ (instead of direct prediction of P) makes the rectification network training smooth and stable.

4.3.4 Experiments

Datasets and Metrics

This section describes a list of datasets and evaluation metrics that are used in the experiments.

Datasets: All ESIR models to be evaluated are trained by using the Synth90K and SynthText, and there is no fine-tuning by using any third dataset. The ESIR models are evaluated over 6 scene text datasets including 3 normal datasets ICDAR2013, IIIT5K and SVT where most scene texts are almost horizontal, and 3 distorted datasets ICDAR2015, SVTP and CUTE80 where a large amount of scene texts suffer from perspective and curvature distortions. The 6 datasets are publicly accessible which have been widely used for evaluations in scene text recognition research.

Synth90K [6] contains 9 million synthetic text images with a lexicon of 90K, and it has been widely used for training scene text recognition models. It has no separation of training and test data and all images are used for training.

SynthText [5] is the synthetic image dataset that was created for scene text detection research. It has been widely used for scene text recognition research as well by cropping text image patches using the provided annotation boxes. State-of-the-art methods crop different amounts for evaluations, e.g. [191] crops 4 million, Shi [190] crops over 7 million, etc. We crop 4 million text image patches from this dataset which are at lower end for fair benchmarking.

ICDAR2013 [159] is used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2013. It contains 848 images for model training and 1095 for testing.

ICDAR2015 [163] was used in the Robust Reading Competition under ICDAR 2015. It contains incidental scene text images that are captured without preparation before capturing. 2077 text image patches are cropped from this dataset, where a large amount of cropped scene texts suffer from perspective and curvature distortions.

IIIT5K [165] consists of 2000 training images and 3000 test images that are cropped from scene texts and born-digital images. Each word image in this dataset has a 50-word lexicon and a 1000-word lexicon, where each lexicon consists of a ground-truth word and a set of randomly picked words.

SVT [166] is collected from the Google Street View images that were used for scene text detection research. 647 words images are cropped from 249 street view images and words within most cropped word images are almost horizontal. Each word image has a 50-word lexicon.

SVTP [184] consists of 639 word images that are cropped from the SVT images. Most images in this dataset are heavily distorted by perspective distortions which are specifically picked for evaluation of scene text recognition under perspective views. Each word image has a 50-word lexicon as inherited from the SVT dataset.

CUTE [185] consists of 288 word images where most cropped scene texts are curved. All word images are cropped from the CUTE dataset which contains 80 high-resolution scene text images that are originally collected for the scene text detection research. No lexicon is provided for the 288 images in this dataset.

Metrics: We follow the protocol and evaluation metrics that have been widely used in scene text recognition research [190, 227]. In particular, the recognition covers 68 characters including 10 digits, lower-case letters and 32 ASCII punctuation marks. In evaluation, only digits and letters are counted and the rest is directly discarded. If a lexicon is provided, the lexicon word that has the minimum edit distance with the predicted word is selected. In addition, evaluations are based on the correctly recognized words (CRW) which can be determined based on the ground truth transcription.

Implementation: The proposed scene text recognition network is implemented by using the Tensorflow framework. The ADADELTA is adopted as optimizer which employs adaptive learning rate and weighted cross-entropy in sequence loss calculation. The network is trained in 1 million iterations with a batch size of 64. In addition, the network training is performed on a workstation with one Intel Core i7-7700K CPU, one NVIDIA GeForce GTX 1080 Ti graphics card with 12GB memory and 32GB RAM.

Three ESIR models are trained for evaluations and benchmarking with state-of-the-art techniques. The first is a baseline model **ESIR [VGG, SK]** as shown in Table 4.1, which uses VGG as the network backbone and the Synth90 as the training data. The second model as denoted by **ESIR [ResNet, SK]** uses the same training data but ResNet as the network backbone. The third model as denoted by **ESIR [ResNet, SK+ST]** uses ResNet as the network backbone but a combination of

TABLE 4.1: Scene text recognition performance over the datasets ICDAR2013 (IC13), ICDAR2015 (IC15), IIIT5K, SVT, SVTP and CUTE, where "50" and "1K" in the second row denote the lexicon size and "None" means no lexicon used. The used network backbone and training data are given in [] at the end of each method, where "SK" and "ST" denote the Synth90K and SynthText datasets, respectively.

Methods	IC13	IC15	IIIT5K			SVT		SVTP	CUTE
	None	None	50	1k	None	50	None	None	None
Jaderberg [134] [VGG, SK]	81.8	-	95.5	89.6	-	93.2	71.7	-	-
Jaderberg [170] [VGG, SK]	90.8	-	97.1	92.7	-	95.4	80.7	-	-
Shi [162] [VGG, SK]	88.6	-	96.2	93.8	81.9	95.5	81.9	71.8	59.2
Yang [228] [VGG, Private]	-	-	97.8	96.1	-	95.2	-	75.8	69.3
Cheng [227] [ResNet, SK+ST]	**93.3**	70.6	99.3	97.5	87.4	97.1	85.9	71.5	63.9
Cheng [191] [VGG, SK+ST]	-	68.2	99.6	98.1	87.0	96.0	82.8	73.0	76.8
Shi [190] [ResNet, SK+ST]	91.8	76.1	**99.6**	**98.8**	**93.4**	97.4	89.5	78.5	79.5
ESIR [VGG, SK]	87.4	68.4	95.8	92.9	81.3	96.7	84.5	73.8	68.4
ESIR [ResNet, SK]	89.1	70.1	97.8	96.1	82.9	97.1	85.9	75.8	72.1
ESIR [ResNet, SK+ST]	91.3	**76.9**	**99.6**	**98.8**	93.3	**97.4**	**90.2**	**79.6**	**83.3**

the Synth90K and SynthText as training data, largely for benchmarking with state-of-the-art models such as ASTER and AON that also use a combination of the two datasets in training. All three ESIR models are trained under the same parameters setting: rectification iteration number: 5; number of line segments: 20; order of the middle line polynomial: 4.

Experimental Results

Rectification and Recognition: The proposed ESIR has been evaluated extensively over the 6 public datasets as described in **Dataset** that contain both normal scene text images and scene text images with a variety of perspective and curvature distortions. In addition, it has been compared with a number of state-of-the-art scene text recognition techniques that employ rectification, feature learning techniques, etc. as described in **Related Work**. Table 4.1 shows experimental results.

As Table 4.1 shows, the **ESIR [ResNet, SK]** consistently outperforms the **ESIR [VGG, SK]** across all 6 datasets evaluated due to the use of a more powerful network backbone. In addition, the **ESIR [ResNET, SK+ST]** consistently outperforms the **ESIR [VGG, SK]** across the six datasets, demonstrating the value of including more data in network training. By taking a second look at the datasets, we observe that Synth90K mainly consists of English words but few sample images with numbers and punctuation, whereas SynthText contains a large amount

of sample images with numbers and punctuation. This could partially explain why the inclusion of the SynthText helps more for the datasets IIT5K and CUTE that contain large amounts of images with numbers and punctuation.

The proposed ESIR achieves superior scene text recognition performance across the 6 datasets as compared with state-of-the-art techniques. For the three distorted datasets, ESIR outperforms state-of-the-art techniques under all different settings, demonstrating the advantage of the proposed iterative rectification network. In particular, **ESIR [VGG, SK]** consistently outperforms the [162] over the SVTP and CUTE when similar network backbone and training data are used. **ESIR [ResNET, SK+ST]** also outperforms [227] and [190] consistently across the ICDAR2015, SVTP and CUTE under the same setting. For the three normal datasets, ESIR also achieves state-of-the-art performance. In particular, both **ESIR [VGG, SK]** and **ESIR [ResNET, SK+ST]** outperform state-of-the-art techniques over the SVT dataset that contains a large amount of low-quality scene texts from street view imagery. For the datasets ICDAR2013, [227] achieves the best accuracy but it requires character-level bounding box annotations. [170] and [162] outperform the **ESIR [VGG,SK]** slightly on the ICDAR2013 under the same setting but they only recognize words within a 90K dictionary. [190] achieves the best accuracy on the IIIT5K, but it crops 7.2 million training images from the SynthText whereas we only crop 4 million training images.

Fig. 4.4 illustrates the scene text rectification and recognition by the proposed ESIR, where the three columns show several sample images from the CUTE and SVTP, the rectified images by using the **ESIR [VGG, SK]**, and the recognized texts without (at the top) and with (at the bottom) the proposed iterative rectification (incorrectly recognized texts are highlighted in red color), respectively. As Fig. 4.4 shows, the proposed ESIR is capable of rectifying scene text images with various perspective and curvature distortions in most cases. For the last two severely distorted scene text images, the rectification could be further improved by employing a larger number of rectification iterations (beyond default 5 under the current setting). At the same time, we can see that the proposed ESIR does not degrade scene text images that do not suffer from perspective and curvature distortions as illustrated in the sixth sample image. Further, the proposed ESIR helps to improve the scene text recognition performance greatly as shown in the third column. Note that the recognition here does not use any lexicon, and the

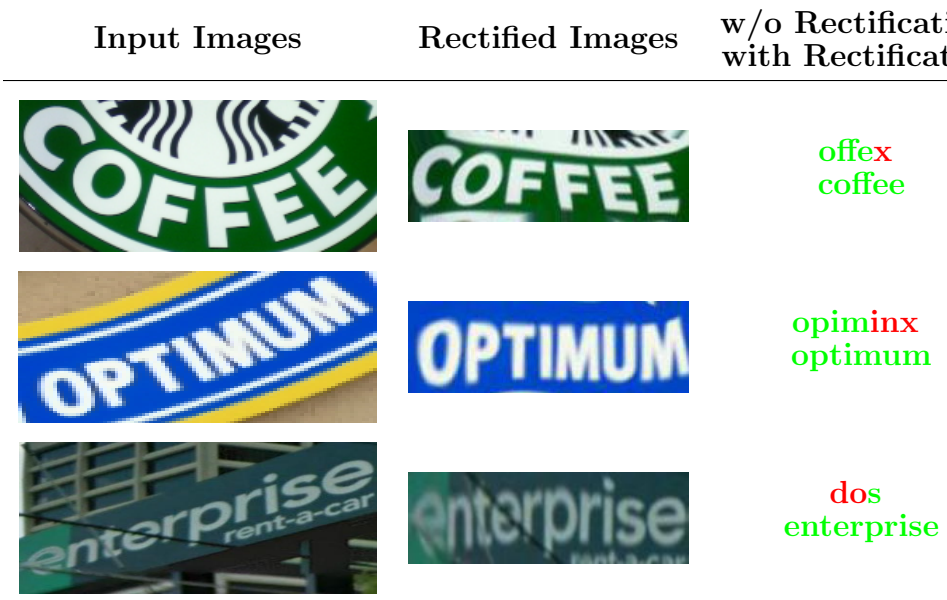


FIGURE 4.4: Illustration of scene text rectification and recognition by the proposed ESIR: For the distorted scene text images from the SVTP and CUTE80 in the first column, the second column shows the finally restored scene text images by the proposed rectification network and the third column shows the recognized texts with and without using the proposed rectification technique, respectively. It is clear that the proposed rectification helps to improve the scene text recognition performance greatly.

recognition performance can be greatly improved by including a lexicon or even a large dictionary, e.g. the mis-recognized 'bookstorf' and 'tacation' from the last two sample images could be corrected if a dictionary is used.

We conjecture that the ESIR's superior recognition performance especially over the three distorted datasets is largely due to our proposed iterative rectification network. Fig. 4.5 compares scene text rectifications by our proposed rectification network and two state-of-the-art rectification networks in RARE [162] and ASTER [190]. As Fig. 4.5 shows, the proposed network produces clearly better rectifications as compared with rectifications by RARE and ASTER. The better rectifications are largely due to the robust line-fitting transformation as well as the iterative rectification framework as described in **Proposed Method**.

Ablation Analysis: The performance of the proposed scene text rectification and recognition technique is heavily affected by two key parameters, namely, the number of rectification iterations and the number of line segments as illustrated in Fig. 4.2. We study these two key parameters separately by using the Synth90K and SynthText as training data consistently. Table 4.2 shows experimental results,



FIGURE 4.5: Visual comparison of different scene text rectification methods: For the four sample images in the first column, columns 2-4 show the rectified images by using the RARE, ASTER, and ESIR, respectively. The sample images are from SVTP and CUTE80 which suffer from perspective and curvature distortions as well as complex image background. The proposed ESIR performs clearly better in scene text distortion rectification.

where the first set of experiments fix the number of line segments at 20 but change the rectification iteration number from 1 to 5 and the second set of experiments fix the rectification iteration number at 5 but use 5, 10 and 15 line segments, respectively.

As Table 4.2 shows, the model performance improves consistently when a larger number of rectification iterations is implemented. In particular, the improvement is more significant at the early stage when the first and second iterations of rectification are implemented i.e. the number of iterations changes from 0 to 1 and from 1 to 2. This can be observed more clearly over the two highly distorted datasets SVTP and CUTE as shown in Table 4.2. In addition, using a larger number of line segments also helps to improve the scene text recognition performance though the improvement is not as significant as implementing a larger number of rectification iterations. We conjecture that using a larger number of line segments helps to produce better estimations of text line poses which further helps to improve the scene text rectification and recognition performance.

Computational Costs: Though the proposed ESIR performs multiple iterations of rectification, the overall computation costs just increase slightly as compared with state-of-the-art rectification techniques without using iterations [162, 190]. In

TABLE 4.2: ESIR recognition accuracy under different parameter settings: N denotes the rectification iteration number, L denotes the number of line segments, and default setting uses 5 rectification iterations and 20 line segments. When N or L changes, all other parameters keep the same as the default setting (Synth90K and SynthText are used in training under all settings).

Methods	ICDAR2015	SVTP	CUTE
Default: N=5, L=20	**76.9**	**79.6**	**83.3**
N = 0, L=20	73.9	73.2	73.4
N = 1, L=20	75.8	77.3	78.8
N = 2, L=20	76.3	78.7	81.1
N = 3, L=20	76.7	79.3	82.7
N = 4, L=20	**76.9**	79.5	83.1
N=5, L = 5	75.8	78.0	81.7
N=5, L = 10	76.6	78.9	82.6
N=5, L = 15	**76.9**	79.3	83.0

particular, the proposed ESIR with 5 rectification iterations takes 3ms per image in training with a batch-size of 64 and 28ms per image in testing with a batch-size of 1. Under the similar network setting, the ASTER takes 2.4ms per image in training and 20ms per image in testing. The similar computational cost is largely due to the proposed rectification network which is small and computational light as compared with the feature extraction network and the sequence recognition network.

4.4 Geometry-Aware Image Translation

We designed an innovative Geometry-Aware Domain Adaptation Network (GA-DAN) [36] [4], an end-to-end trainable network that learns and models domain shifts in appearance and geometry spaces simultaneously as illustrated in the last two rows of Fig. 4.6. One unique feature of the proposed GA-DAN is a multi-modal spatial learning structure that learns multiple spatial transformations and converts a source-domain image into multiple target-domain images realistically as illustrated in Figs. 4.9 and 4.10. In addition, a novel disentangled cycle-consistency loss is designed which guides the GA-DAN learning towards optimal transfer concurrently in both geometry and appearance spaces. The GA-DAN takes the cycle structure as illustrated in Fig. 4.7, where the spatial modules (S_X and S_Y) model the feature shifts in geometry space and the generators (G_X and G_Y) complete the

[4]The work has been published in ICCV 2019

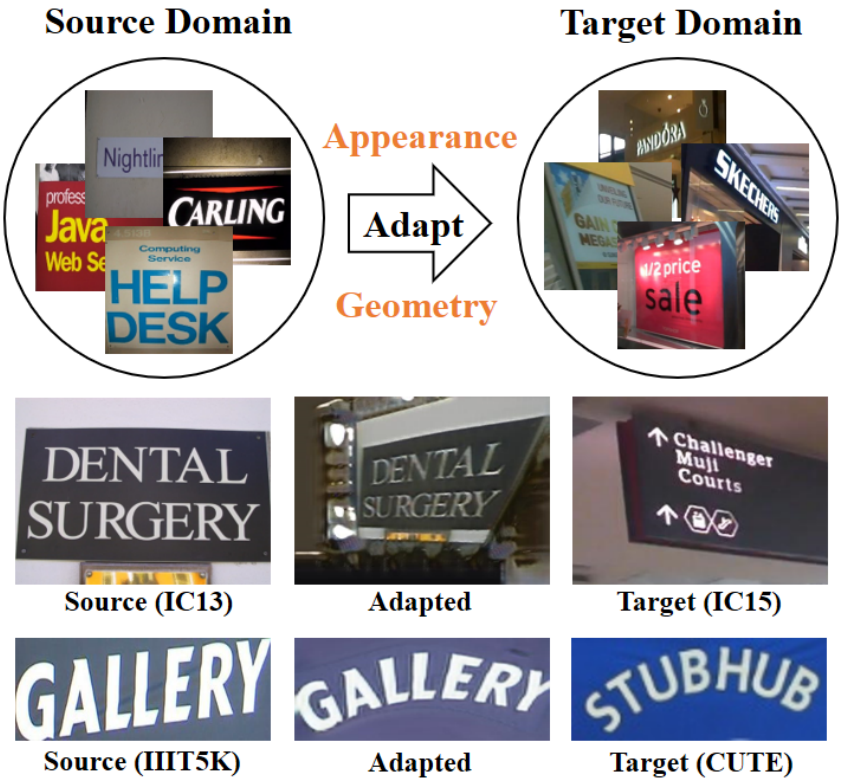


FIGURE 4.6: Domain adaptation by the proposed GA-DAN: For scene text images with clear shifts from the **Source Domain** to the **Target Domain**, GA-DAN models the domain shifts in appearance and geometry spaces simultaneously and generates **Adapted images** with high-fidelity in both appearance and geometry spaces.

blank as introduced by the spatial transformation and model the feature shifts in appearance space. The discriminators discriminate not only 'fake image' and 'real image' but also 'fake transformation' and 'real transformation', leading to optimal modelling of domain and feature shifts in geometry and appearance spaces.

The contributions of this work are threefold. First, it designs a novel network that models domain shifts in geometry and appearance spaces simultaneously. To the best of our knowledge, this is the first network that performs domain adaptation in both spaces concurrently. Second, it designs an innovative multi-modal spatial learning mechanism and introduces a spatial transformation discriminator to achieve multi-modal adaptation in geometry space. Third, it designs a disentangled cycle-consistency loss that balances the cycle-consistency for concurrent adaptation in appearance and geometry spaces and can also be applied to generic domain adaptation.

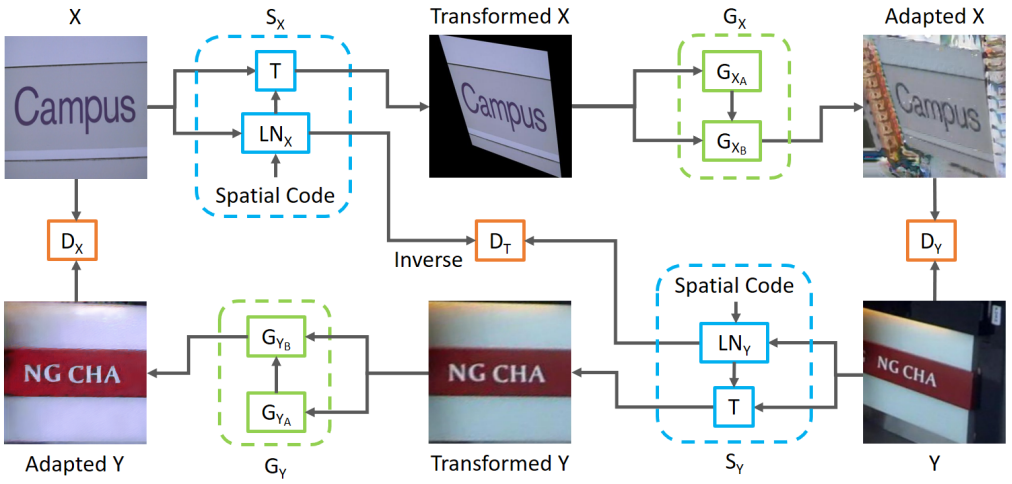


FIGURE 4.7: The structure of the proposed GA-DAN: S_X (or S_Y) represents the spatial modules as enclosed in blue-color boxes which consist of *Spatial Code*, transformation module T and localization network LN_X (or LN_Y) that predict transformation matrix and transform input images. G_X (or G_Y) denote generators consisting of G_{X_A} (or G_{Y_A}) and G_{X_B} (or G_{Y_B}) as enclosed in green-color boxes that complete the background and translate the image style, respectively. D_X, D_Y and D_T within orange-color boxes denote different discriminators.

We propose an innovative geometry-aware domain adaptation network (GA-DAN) that performs multi-modal domain adaptation concurrently in both spatial and appearance spaces as shown in Fig. 4.7. Detailed network architecture, multi-modal spatial learning and adversarial training strategy will be presented in the following three subsections.

4.4.1 Model Architecture

The GA-DAN consists of spatial modules, generators and discriminators as enclosed within blue-color, green-color and orange-color boxes, respectively, as shown in Fig. 4.7. The overall network is designed in a cycle structure, where the mappings between the source domain X and the target domain Y are learned by sub-modules $X \rightarrow$ *Adapted* X ($X \rightarrow Y$) and $Y \rightarrow$ *Adapted* Y ($Y \rightarrow X$), respectively. In the $X \rightarrow Y$ mapping, the spatial module S_X transforms images in X to new images in *Transformed* X that has similar spatial styles as Y. The generator G_X then completes the blank as introduced by the spatial transformation and translates the completed images to new images in *Adapted* X that has similar appearance as Y. A discriminator D_Y attempts to distinguish *Adapted* X and Y which drives S_X

and G_X to learn better spatial and appearance mapping from X to Y. Similar processes happen in the $Y \rightarrow X$ mapping as well.

The spatial modules S_X (as well as S_Y) has a localization network LN_X and a transformation module T for domain adaptation in geometry space, more details to be presented in the following subsection. The generator G_X (as well as G_Y) consists of two sub-generators G_{X_A} and G_{X_B} for adaptation in appearance space. In particular, the spatial module S_X will produce a binary map with 1 denoting the pixels transformed from the original image and 0 for padded black background (not shown but can be inferred from the sample image in Transformed X in Fig. 4.7). Under the guidance of the binary map, G_{X_A} will learn from Y for new contents to complete the black background of the transformed image (as in *Transformed* X), and G_{X_B} further adapts the completed images to have similar appearance as Y as illustrated in Fig. 4.7. Our study shows that the *Adapted* X is quite blurry if a single generator is used to complete the black background and adapt the appearance. The use of the two dedicated generators G_{X_A} and G_{X_B} for background completing and appearance adaptation helps greatly for realistic adaptation in appearance space.

Note directly concatenating an appearance-transfer GAN (e.g. CycleGAN [113]) and a geometry-transfer GAN (e.g. ST-GAN [1]) does not perform well for simultaneous image adaptation in geometry and appearance spaces. Due to the co-existence of spatial and appearance shifts between the source and target domain images, the discriminator of the geometry-transfer GAN (or appearance-transfer GAN) will be confused by the appearance (or geometry) shift which leads to poor adversarial learning outcome. Our GA-DAN is an end-to-end trainable network that coordinates the learning in geometry and appearance spaces simultaneously and drives the network for optimal adaptation in both spaces.

4.4.2 Multi-Modal Spatial Learning

To generate images with different spatial views and features (similar to images in the target domain), we design a multi-modal spatial learning structure that learns multi-modal spatial transformations and maps a source-domain image to multiple target-domain images with different spatial views. Specifically, the multi-modal

spatial learning first samples *Spatial Code* (i.e., random vectors) from normal distributions and then regresses it to predict spatial transformation matrix (according to the spatial features of the input image) by using a localization network LN_X (or LN_Y). With the predicted transformation matrix that could be affine, homography or thin plate spline [183], the input image can be transformed to a new image with a different spatial view by T which performs actual transformation. Multiple new spatial views can be generated by running GA-DAN and sampling the *Spatial Code* multiple times.

The multi-modal spatial learning as guided by D_X and D_Y tends to be unstable and hard to converge as the concurrent learning in geometry and appearance spaces is over-flexible and often entangled with each other. We address this issue by including a new discriminator D_T as shown in Fig. 4.7 which imposes certain constraints to the cyclic spatial learning and accordingly leads to more stable and efficient learning of the whole network. As shown in Fig. 4.7, S_X predicts a transformation matrix H_{XY} for mapping from domain X to domain Y, and S_Y predicts another transformation matrix H_{YX} for mapping from domain Y to domain X. The inverse matrix H_{XY}^{-1} can be derived from H_{XY} and it should be in the same transformation domain with the H_{YX}. D_T thus attempts to discriminates H_{XY}^{-1} and H_{YX} which drives the spatial transformations in two inverse directions to learn from each other. It bridges the spatial learning in opposite directions and imposes extra constraints in the geometry spaces, greatly improving the learning efficiency and learning stability of the whole network.

4.4.3 Adversarial Training

Due to the adaptation in geometry and appearance spaces, the adversarial learning needs to coordinate the minimization of cycle-consistency loss in both spaces properly. In addition, the adversarial learning also needs to take care of the new discriminator D_T as shown in Fig. 4.7. We design an innovative disentangled cycle-consistency loss and adversarial objective to tackle these challenges, more details to be described in the following two subsections.

Disentangled Cycle-Consistency Loss. We design a disentangled cycle-consistency loss that decomposes the cycle-consistency loss into a spatial cycle-consistency loss (SCL) and an appearance cycle-consistency loss (ACL) and balances their weights

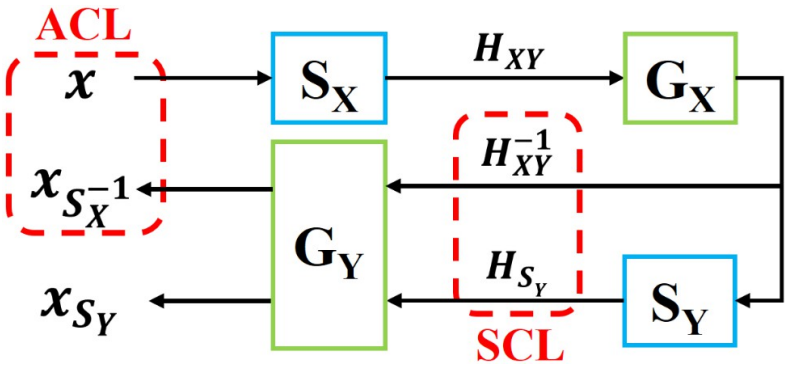


FIGURE 4.8: Illustration of the disentangled cycle-consistency loss: S_X, S_Y, G_X and G_Y denotes the spatial modules and generators, respectively, as shown in Fig. 4.7. x, $x_{S_X^{-1}}$ and x_{S_Y} refer to the input images in domain X, reconstructed image by the inverse transformation of S_X and reconstructed image by S_Y. H_{XY} and H_{S_Y} refer to the predicted transformation matrices by S_X and S_Y. ACL and SCL denote appearance cycle-consistency loss and spatial cycle-consistency loss which are obtained by calculating the L1 loss between $(x, x_{S_X^{-1}})$ and (H_{XY}^{-1}, H_{S_Y}), respectively.

during learning. With spatial transformation involved, a small shift (due to inaccurate prediction of the spatial transformation matrix) in geometry space will lead to a very large cycle-consistency loss which can easily override the ACL and ruin the learning of the whole network. The decomposition of the cycle-consistency loss into ACL and SCL helps to address this issue effectively.

As shown in Fig. 4.8, the image x is fed into S_X to predict the transformation matrix H_{XY} and the transformed image is then fed to G_X for translation in appearance space. The translated image will be recovered in two different manners. First, it will be transformed by the inverse of H_{XY} (i.e. H_{XY}^{-1}) and further translated by G_Y to generate $x_{S_X^{-1}}$. Second, it will be passed to S_Y to predict the transformation matrix H_{S_Y} so as to be transformed by the estimated H_{S_Y} and further translated by G_Y to produce x_{S_Y}. Note the *Spatial Code* in S_X and S_Y are identical here so that the input image can be recovered in geometry space.

The $x_{S_X^{-1}}$ can be fully recovered from x in geometry space since the recovering matrix H_{XY}^{-1} is simply the inverse of H_{XY}. But $x_{S_X^{-1}}$ is different from x in appearance space. The ACL can thus be computed by L1 loss between x and $x_{S_X^{-1}}$ (only appearance difference exists) as follows:

$$ACL_X = E_{x \sim X}[\left\| x_{S_X^{-1}} - x \right\|] \tag{4.1}$$

Though $x_{S_X^{-1}}$ and x_{S_Y} differ only in geometry space, SCL cannot be obtained by computing L1 loss between them because a minor shift in geometry space will lead to a very large L1 loss. To ensure the spatial cycle-consistency, we obtain the SCL by directly computing the L1 loss of the transformation matrix H_{XY}^{-1} and H_{S_Y} as follows:

$$SCL = E_{x\sim X}[\left\| H_{XY}^{-1} - H_{S_Y} \right\|] \tag{4.2}$$

Further, the bordering regions of the original image may be lost by the spatial transformation which could affect the training seriously. While adapting an image from domain X to domain Y, the adaptation should ensure that all image information within the domain X is well preserved. Given the binary transformation map m from S_X, we can directly apply the inverse transformation H_{XY}^{-1} to m to obtain $m_{H_{XY}^{-1}}$. As the missing region by the spatial transformation will not be recovered, a region missing loss (RML) is defined for better preserving the transformed image as follows:

$$RML = E_{x\sim X}[\left\| m_{H_{XY}^{-1}} - m \right\|] \tag{4.3}$$

The overall cycle-consistency loss in the domain X can thus be formed as follows:

$$L_{cyc} = \lambda_{acl} ACL + \lambda_{scl} SCL + RML \tag{4.4}$$

where λ_{acl} and λ_{scl} are the weights of ACL and SCL.

Adversarial Objective. The adversarial objective of the mapping $X \to Y$ can be defined by:

$$\begin{aligned} L_{GAN} &= E_{y\sim Y}[\log D_Y(y)] + E_{x\sim X}[\log(1 - D_Y(G_X(S_X(x))))] \\ &+ E_{y\sim Y}[\log D_T(H_{YX}^{-1})] + E_{x\sim X}[\log(1 - D_T(H_{XY}))] \end{aligned} \tag{4.5}$$

where H_{XY} and H_{YX}^{-1} are the transformation matrix for $X \to Y$ and the inverse transformation for $Y \to X$. S_X and G_X aim to minimize this objective while D_Y and D_T try to maximize it, i.e. $minS_X G_X maxD_Y D_T L_{GAN}$. The objective of the mapping $Y \to X$ can be obtained similarly. Note to ensure that the translated

image preserves features of the original image, an identity loss is included as follows:

$$L_{idt} = E_{x \sim X}[\|G_X(S_X(x)) * m - S_X(x) * m\|] \tag{4.6}$$

where m refers to the binary mask as produced by S_X.

4.4.4 Experiments

The proposed image adaptation technique has been evaluated over the scene text detection and recognition tasks.

Datasets: The experiments involve seven publicly available scene text detection and recognition datasets as listed:

ICDAR2013 [159] is used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2013. The images explicitly focused around the text content of interest. It contains 848 images for network training and 1095 for testing.

ICDAR2015 [163] is used in the Robust Reading Competition under ICDAR2015. It contains incidental scene text images that appears in the scene without taking any specific prior action to improve its positioning / quality in the frame.

MSRA-TD500 [229] dataset consists of 500 natural images (300 for training, 200 for test), which are taken from indoor and outdoor scenes using a pocket camera. The indoor images mainly capture signs, door plates and caution plates while the outdoor images mostly capture guide boards and billboards with complex background.

IIIT5K [165] has 2000 training images and 3000 test images that are cropped from scene texts and born-digital images. Each word in this dataset has a 50-word lexicon and a 1000-word lexicon, where each lexicon consists of a ground-truth word and a set of randomly picked words.

SVT [166] is collected from the Google Street View images that were used for scene text detection research. 647 words images are cropped from 249 street view images and most cropped texts are almost horizontal.

TABLE 4.3: Scene text detection over the test images of the target datasets ICDAR2015 and MSRA-TD500: 'IC13', 'Target', 'AD-IC13' and '10-AD-IC13' denote the dataset ICDAR2013, target dataset (ICDAR2015 or MSRA-TD500), 1-to-1 adapted ICDAR2013 and 1-to-10 adapted ICDAR2013, respectively. 'SynthText' refers to 800K synthetic images as reported in [5].

Methods	ICDAR2015			MSRA-TD500		
	Recall	**Precision**	**F-score**	**Recall**	**Precision**	**F-score**
RRD [230] [SynthText + Target]	79.0	85.6	82.2	73.0	**87.0**	**79.0**
TextSnake [231] [SynthText + Target]	80.4	84.9	82.6	**73.9**	83.2	78.3
EAST [IC13]	43.7	68.2	53.3	34.9	71.2	46.8
EAST [AD-IC13]	59.6	69.9	64.4	51.5	67.7	58.5
EAST [10-AD-IC13]	71.6	67.3	69.4	55.8	69.9	62.1
EAST [Target]	76.9	81.1	79.0	64.4	73.8	68.7
EAST [IC13 + Target]	77.0	83.2	80.0	66.2	74.8	70.3
EAST [AD-IC13 + Target]	79.2	83.7	81.4	67.7	77.5	72.3
EAST [10-AD-IC13 + Target]	**81.6**	**85.6**	**83.5**	71.1	80.5	75.5

SVTP [184] has 639 word images that are cropped from the SVT images. Most images in this dataset suffer from perspective distortion which are purposely selected for evaluation of scene text recognition under perspective views.

CUTE [185] has 288 word images most of which are curved. All words are cropped from the CUTE dataset which contains 80 scene text images that are originally collected for scene text detection research.

Scene Text Detection: The proposed GA-DAN is evaluated by the performance of the scene text detectors that are trained by using its adapted images. In evaluations, the training set of ICDAR2013 (IC13) is used as the source dataset and the training sets of ICDAR2015 (IC15) and MSRA-TD500 (MT) are used as the target datasets which contain very different images as compared with those in IC13. GA-DAN generates two sets of images 'AD-IC13' and '10-AD-IC13' for each of the two target datasets. The 'AD-IC13' is generated by 1-to-1 adaptation where each IC13 image is transformed to a single image that has similar geometry and appearance as the target dataset. The '10-AD-IC13' is produced by 1-to-10 adaptation where each IC13 image is transformed to 10 adapted images by sampling 10 different spatial codes. Scene text detector EAST [160] is adopted for evaluation.

Table 4.3 shows quantitative results on the test set of two target datasets. Seven EAST models are trained for each target dataset by using different training images including 1) [IC13]: the training set of IC13, 2) [AD-IC13]: the 1-to-1 adapted IC13, 3) [10-AD-IC13]: the 1-to-10 adapted IC13, 4) [Target]: the training set of each

target dataset, 5) [IC13 + Target]: the combination of the IC13 training set and the training set of each target dataset, 6) [AD-IC13 + Target]: the combination of AD-IC13 and the training set of each target dataset, and 7) [10-AD-IC13 + Target]: the combination of 10-AD-IC13 and the training set of each target dataset.

As Table 4.3 shows, the effectiveness of GA-DAN adapted images can be observed from three aspects. First, **EAST [AD-IC13]** outperforms **EAST [IC13]** by f-scores of 11.1% (53.3% → 64.4%) and 11.7% (46.8% → 58.5%) on the target datasets IC15 and MT, respectively, demonstrating the effectiveness of GA-DAN in adapting images from IC13 to IC15 and MT. Second, **EAST [10-AD-IC13]** improves **EAST [AD-IC13]** by f-scores of 5% (64.4% → 69.4%) and 3.6% (58.5% → 62.1%) on IC15 and MT, respectively. This shows the effectiveness of the multi-modal spatial learning that transforms a source-domain image to multiple target-domain images that are complementary with different spatial views. Third, **EAST [10-AD-IC13+Target]** improves **EAST [IC13+Target]** by a f-score of 3.5% (80.0% → 83.5%) and 5.2% (70.3% → 75.5%) on IC15 and MT, respectively. This shows that the adapted images are clearly more useful when combined with the training images of the target datasets for model training.

In addition, **EAST [10-AD-IC13+Target]** achieves state-of-the-art performance on the dataset IC15 by including only 2.3K GA-DAN adapted images (from 230 IC13 training images). As a comparison, TextSnake [231] uses 800K synthetic images in 'SynthText' [232] and it is also more advanced scene text detectors. Though the '10-AD-IC13' is much smaller than SynthText, it contributes more to the detection improvement largely because of the large domain shifts between SynthText and IC15. For the target dataset MT, the f-score of **EAST [10-AD-IC13+Target]** is slightly lower than that of state-of-the-art detectors TextSnake and RRD, largely because the domain shifts between MT and SynthText are relatively small and the much larger amount of images in SynthText help more on the performance improvement. We believe higher f-score can be achieved when a higher number of GA-DAN adapted images are included in model training.

Table 4.4 shows the detection performance of different domain adaptation methods when EAST are trained by using their adapted images from IC13 to IC15 (the Baseline is trained using the original IC13 training images). Note for CycleGAN we adopt patch-wise training to minimize the effect of geometry differences in adversarial training. As ST-GAN is originally for image composition, we adapt

TABLE 4.4: Scene text detection on the IC15 test images: The detection models are trained using the adapted IC13 training images (from IC13 to IC15) by different adaptation methods as listed. (Baseline is trained by using the original IC13 training images)

Methods	Recall	Precision	F-score
CycleGAN [113]	50.3	66.3	57.2
ST-GAN [1]	52.9	63.4	57.6
ST-GAN + CycleGAN	57.3	64.7	60.8
Baseline	43.7	68.2	53.3
GA-DAN	**59.6**	**69.9**	**64.4**

it to achieve image translation in geometry space and restrict the transformation parameters to avoid boundary losing in testing phase. As Table 4.4 shows, all three adaptation models GA-DAN, CycleGAN and ST-GAN outperform Baseline clearly, and GA-DAN achieves clearly better f-score (64.4% vs. 57.2% and 57.6%), demonstrating its superiority in adapting more realistic images. We also evaluate a new model ST-GAN + CycleGAN that directly concatenates ST-GAN and CycleGAN for adaptation in both geometry and appearance spaces. It shows that our GA-DAN still performs better by a large margin in f-score (64.4% vs. 60.8%), demonstrating its advantages in concurrent learning of geometric and appearance features.

Fig. 4.9 compares our GA-DAN with several state-of-the-art image adaptation methods. As Fig. 4.9 shows, GA-DAN adapts in both appearance and geometric spaces realistically, whereas SimGAN and CycleGAN can only adapt appearance features and ST-GAN can only adapt geometric features. In addition, GA-DAN_1, GA-DAN_2 and GA-DAN_3 show three GA-DAN adapted images with different spatial views, demonstrating the effectiveness of our proposed multi-modal spatial learning.

Scene Text Recognition: For scene text recognition experiment, we select the CUTE [185] and SVTP [184] as the target datasets. As current scene text recognition datasets are all too small, we combine all images from datasets IC13 [159], IIIT5K [165] and SVT [166] as the source dataset denoted by 'COMB'. As scene texts in CUTE and SVTP are most curved or in perspective views but most COMB texts are horizontal, we use the thin plate spline for spatial transformation which is flexible for various spatial transformations.



FIGURE 4.9: Comparing our GA-DAN with state-of-the-art adaptation methods: The first and last columns show source-domain (IC13) and target-domain (IC15) images. GA-DAN_1, GA-DAN_2 and GA-DAN_3 show three GA-DAN adapted images of different spatial views.

Table 4.5 shows recognition accuracy when COMB images are adapted by different adaptation methods and then used to train the scene text recognition model: MORAN [233]. As Table 4.5 shows, **GA-DAN [WM]** (GA-DAN with 1-to-1 spatial transformation) outperforms other adaptation methods with a large margin. Additionally, most compared adaptation methods do not show clear improvement over the Baseline (trained by using the original COMB images without adaptation). In particular, CycleGAN and CyCADA improve the accuracy by 1.0% and 1.3% only for CUTE because they only adapt in appearance space but the main discrepancy between COMB and CUTE is in geometry space. CoGAN and UNIT tend to over-adapt the text appearance which may even change the text semantics and make texts unrecognizable.

Table 4.5 also shows the ablation study results. Two GA-DAN models are trained for image adaptation. The first model is a complete GA-DAN with all newly designed features and components includes. The second is **GA-DAN [WD]** which is trained with a normal instead of disentangled cycle-consistency loss. For fair comparison, the region missing loss is also included in **GA-DAN [WD]**. For the complete GA-DAN, three sets of adapted images are generated to train the recognition model. The first set is **GA-DAN [WA]** that just takes the output of S_X without appearance adaptation as shown in Fig. 4.7. The second set is **GA-DAN [WM]** that performs 1-to-1 adaptation and transforms each source-domain image into a single target-domain image. The third set is **GA-DAN [10 AD]** that performs 1-to-10 adaptation and transforms each source-domain image

TABLE 4.5: Ablation study and comparisons with state-of-the-art adaptation methods: Recognition models are trained by different adaptations of the COMB images to the target domains CUTE and SVTP (Baseline uses the original COMB images and 'Random' applies random spatial transformation in adaptation).

Methods	COMB-CUTE	COMB-SVTP
SimGAN [22]	30.7	42.6
UNIT [182]	28.7	40.8
CoGAN [234]	28.3	40.2
DualGAN [224]	31.5	42.7
CycleGAN [113]	31.9	43.0
CyCADA [181]	32.2	43.6
Baseline	30.9	42.5
Random	28.8	42.7
GA-DAN [WD]	32.6	44.9
GA-DAN [WA]	36.1	45.2
GA-DAN [WM]	38.2	47.1
GA-DAN [10 AD]	**43.1**	**51.7**



FIGURE 4.10: Comparing GA-DAN with state-of-the-art adaptation methods: Rows 1-2 show adaptation from COMB to CUTE, Rows 3-4 show adaptation from COMB to SVTP. GA-DAN_1, GA-DAN_2 and GA-DAN_3 show GA-DAN adapted images of different spatial views.

into 10 target-domain images. As Table 4.5 shows, **GA-DAN [WA]** clearly outperforms Baseline and 'Random' (adapted using a random spatial transformation matrix) as well as state-of-the-art adaptation methods, showing the superiority of our spatial module in learning correct and accurate spatial transformations. **GA-DAN [WD]** outperforms state-of-the-art methods but clearly performs worse than **GA-DAN [WM]**, demonstrating the effectiveness of the proposed disentangled cycle-consistency loss. **GA-DAN [10 AD]** outperforms **GA-DAN [WM]** clearly, demonstrating the effectiveness of our proposed multi-modal spatial learning.

Fig. 4.10 compares our GA-DAN with several state-of-the-art adaptation methods.

As Fig. 4.10 shows, GA-DAN adapts in both appearance and geometry spaces realistically whereas CycleGAN and SimGAN can only adapt in appearance space. In addition, GA-DAN_1, GA-DAN_2 and GA-DAN_3 show that the proposed GA-DAN is capable of transforming a source-domain image to multiple target-domain images of different spatial views.

4.5 Conclusion

This chapter presents an end-to-end trainable scene text recognition network (ESIR) that is capable of recognizing distorted scene texts via iterative rectification. The proposed network estimates and corrects perspective distortion and text line curvature iteratively as driven by better scene text recognition performance. In particular, a novel line-fitting transformation is designed to estimate the pose of text lines in scenes, and an iterative rectification framework is developed for optimal scene text rectification and recognition. The proposed network is also robust to parameter initialization and does not require extra annotations. Experiments over a number of public datasets demonstrate its superior performance in scene text rectification and recognition.

In addition, we present a geometry-aware domain adaptation network that is capable of achieving domain adaptation in both geometry and appearance spaces simultaneously. A multi-modal spatial learning technique is proposed which can generate multiple adapted images with different spatial views. A novel disentangle cycle-consistency loss is designed which greatly improves the stability and concurrent learning in both geometry and appearance spaces. The proposed network has been validated over scene text detection and recognition tasks and experiments show the superiority of the adapted images while applied to train deep networks.

Chapter 5

3D-Level Image Synthesis

5.1 Introduction

Image synthesis in 2-dimensional (2D) space are significantly constrained by the 2D image views. For example, image composition with 2D foreground objects can only generate new images with a fixed viewpoint for the embedded foreground objects. Image synthesis in 3-dimensional (3D) space can greatly mitigate such constraints by embedding 3D foreground object models in various poses and views. We investigated 3D image synthesis, targeting to embed 3D virtual objects into 2D background images realistically.

To embed 3D virtual objects into 2D background images realistically, we need to recover the environment illumination of the background image so as to re-light/render the 3D virtual objects harmoniously. Lighting estimation aims to recover illumination from a single image with limited field of view. It has a wide range of applications in various computer vision and computer graphics tasks such as high-dynamic-range (HDR) relighting in mixed reality, etc. However, lighting estimation is an under-constrained problem as it aims to recover a 360-degree full-view illumination map from an image with limited field of view. In addition, high-dynamic-range (HDR) illumination is required to be inferred from low-dynamic-range (LDR) observations for the purpose of realistic object relighting.

Lighting estimation has been tackled through direct generation of illumination maps [61, 62, 65] and regression of parameters of representative illumination functions such as spherical harmonics function [2, 59] and spherical Gaussian function [3, 60]. However, the functional representation methods struggle to regress accurate frequency information (especially high-frequency information) that often leads to inaccurate shading and shadow effects in relighting [2] or requires complex optimization steps [3]. Directly generating illumination maps can preserve some high-frequency information, but it can hardly recover other information of the light sources such as light directions and sizes [235].

Further, lighting estimation in frequency domain normally represents illumination with Spherical Harmonics (SH) which lack spatial localization capabilities. Thus, it tends to capture global lighting instead of the exact spatial locations of the light sources which often leads to inaccurate shading and shadow effects for the embedded 3D objects. Lighting estimation in spatial domain has been addressed by direct generation of the illumination maps or indirect reconstruction through spherical Gaussian function. However, direct generation of illumination maps often leads to worse generalization as lighting estimation is an under-constrained problem by itself, and spherical Gaussian often involves a complicated training process as described in [3]. Both types of approaches in spatial domain do not explicitly consider lighting frequency, and thus lead to inaccurate relighting performance. The high frequency information also tends to be blurred due to the use of naive L2 loss in the training. Besides the aforementioned constraints, existing evaluation metrics in lighting estimation usually assess the objects rendered with the predicted illumination maps, which is time-consuming and sensitive to the test setting.

To address the aforementioned issues, we propose EMLight [37] [1] which predicts illumination from a single image via a two-stage manner as described in section 5.3 and NeedleLight [38] [2] which predicts illumination with a novel representation basis as described in section 5.4.

[1] The work has been published in AAAI 2021
[2] The work is under submission

5.2 Related Works

5.2.1 Lighting Estimation

Lighting estimation is a classic challenge in computer vision and computer graphics, and it is critical for realistic relighting in virtual objects insertion [43, 50–55, 58, 236]. Traditional approaches require user intervention or assumptions about the underlying illumination model, scene geometry, etc. For example, Karsch et al. [13] recover parametric 3D lighting from a single image but requires user annotations for initial lighting and geometry estimates. Zhang et al. [237] require a full multi-view 3D reconstruction of scenes. Lombardi and Nishino [57] estimate illumination from an object of known shape with a low-dimensional model. [58] makes use of additional depth information to recover spherical harmonics illumination.

The recent learning-based works aim to estimate lighting from images by regressing representation parameters [3, 59, 60] or generating illumination maps [61, 62]. Garon et al. [2] estimate lighting by predicting spherical harmonic (SH) coefficients from a background image and local patch. Gardner et al. [3] estimate the positions, intensities, and colours of light sources and reconstructs illumination maps with a spherical Gaussian function. On top of it, Li et al. [63] represent illumination maps with multiple spherical Gaussian functions and regresses the corresponding Gaussian parameters for lighting estimation. Gardner et al. [61] generate illumination maps directly with a two-steps training strategy. Song and Funkhouser [62] estimate per-pixel 3D geometry and uses a convolutional network to predict unobserved contents in the environment map. LeGendre et al. [64] regress HDR lighting from LDR images by comparing the rendered sphere with predicted illumination to the ground truth. Srinivasan et al. [65] estimate a 3D volumetric RGB model of a scene and uses standard volume rendering to estimate incident illuminations. Given any illumination map, the framework proposed by Sun et al. [238] is able to achieve relighting on the RGB portrait image taken in an unconstrained environment. Besides, an end-to-end Generative Adversarial Network [146] is proposed to generate shadow without explicitly estimating the illumination map.

The aforementioned works either lose realistic frequency information or produce inaccurate light sources in illumination estimation. In the proposed EMLight [37], we decompose illumination maps into three components and design a regression

network with a spherical mover's loss that estimates the decomposed illumination parameters with accurate spatial information. Using the estimated illumination parameters as guidance, our designed neural projector generates illumination maps accurately with realistic frequency information.

5.2.2 Needlets

Needlets are a new generation of spherical wavelets [239] and have desirable localization capabilities in both spatial and frequency domains. Narcowich et al. [240] first introduces needlets in the Functional Analysis literature, and Baldi et al. [241] further analyses the statistical properties of needlets. Due to the good localization property in the multi-pole and pixel spaces, needlets have been widely applied to the research of Cosmic Microwave Background [242, 243].

The aforementioned works estimate lighting in either frequency domain or spatial domain which is insufficient to capture the complex illumination in real scenes. Additionally, most existing works regress the illumination with a naive L2 loss or its variant which struggles to regress high frequency information and often introduces blurs. In the proposed NeedleLight, we introduce needlet basis for lighting representation which allows regression of illumination in both frequency and spatial domains jointly. A novel spherical transport loss is proposed to achieve stable and accurate regression of needlet coefficients. Details on lighting representation, spherical transport loss are to be presented in the ensuing sections.

5.3 Earth Mover's Light

In this work, we propose **EMLight** (**E**arth **M**over's **Light**) [37] [3], an accurate illumination estimation framework that is capable of locating light sources and recover illumination with realistic frequency simultaneously. EMLight consists of an interconnected regression network and neural projector, where the regression network predicts illumination parameters accurately and the neural projector leverages the estimated illumination parameters to synthesize illumination maps with realistic frequency information. Instead of regressing illumination parameters separately

[3]The work has been published in AAAI 2021

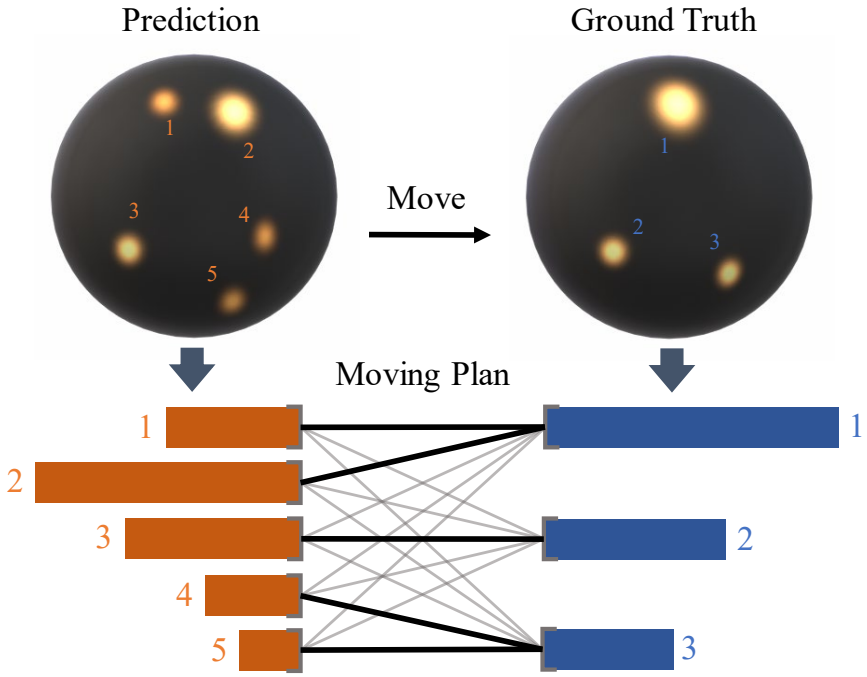


FIGURE 5.1: Illustration of our proposed Earth Mover's Light (EMLight): EMLight treats two illumination maps as two discrete spherical distributions. Motivated by the idea of Earth Mover's distances, we design a spherical mover's loss (SML) to measure the distance between two spherical distributions by calculating the minimum distance of moving one distribution to another along the spherical surface. SML aims to find the best *Moving Plan* (with minimum total distance) as illustrated by connections between the two distributions. The thickness of connecting lines denotes the amount of 'Earth' moved between the two points.

without considering them as a whole as in many existing works, we formulate the overall scene illumination by a spherical distribution and treat the illumination estimation as the regression of a spherical distribution as illustrated in Fig. 5.1.

For accurate illumination representation, we decompose the illumination map into *light distribution*, *light intensity*, and *ambient term* that capture the energy distribution of light sources, the overall intensity of light sources, and the average of remaining energy excluding light sources, respectively. As illumination maps are spherical images, we define N anchor points on a unit sphere to model discrete light distributions. The task of illumination prediction is thus converted to the problem of regressing light distribution, light intensity and ambient term. The light intensity and ambient term are scalar values, which can be directly regressed with a naive L2 loss by the regression network. However, directly regressing N discrete values (at N anchor points) of light distribution with a naive L2 loss or

cross-entropy loss is undesirable as this does not take advantage of the subtleties of spherical distributions such as the spatial information.

Inspired by the Earth Mover's distance [244] that measures the distance between two distributions, we design a spherical mover's loss to regress light distributions by conducting 'Earth Mover' on the unit sphere as illustrated in Fig. 5.1. SML evaluates the distance between two spherical distributions by measuring the minimum radian distance required to move one spherical distribution to another along the spherical surface, and the target is to find the optimal moving plan among all possible moves between two distributions. It captures spatial information of spherical distribution, which greatly helps for accurate estimation of spherical light distribution.

Under the guidance of illumination parameters that are predicted by the regression network, the neural projector generates accurate illumination maps with realistic frequency information in an adversarial manner. Different from normal images, the illumination map is a panorama that usually suffers from different levels of spherical distortions at different latitudes. We therefore adopt spherical convolution [245] for the accurate generation of panoramic illumination maps.

The contribution of this work can be summarized in three aspects. First, we formulate the illumination estimation as the regression of the spherical distribution of illumination. To the best of our knowledge, this is the first work that estimates illumination in this manner. Second, we design a novel spherical mover's loss that takes advantage of the subtleties of spherical illumination distributions. Third, we design a neural projector that employs spherical convolution to synthesize panoramic illumination maps with realistic frequency information through adversarial training.

EMLight consists of two sequential modules including a regression network and a neural projector as illustrated in Figs. 5.2 and 5.3. The illumination parameters estimated by the regression network will guide the neural projector to generate illumination maps accurately.

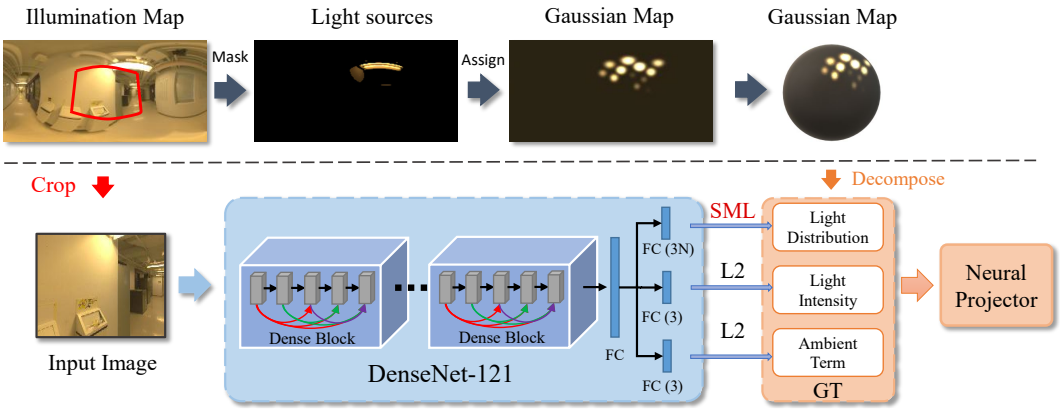


FIGURE 5.2: Illumination decomposition and estimation in EMLight: The upper and lower graphs illustrate the illumination map decomposition and the structure of the light parameter regression network, respectively. Given an *Illumination Map*, we first derive the *Light sources* region via thresholding and then assign light source pixels to N anchor points as illustrated in *Gaussian Map* (visualized by spherical Gaussian function). We decompose the illumination map into light distribution, light intensity and ambient term and use the decomposition as ground truth for regression network training. The regression network employs a local region (as highlighted by the red box) as the input and three fully connected layers (FC) with output size of $3N$, 3 and 3 (RGB images have 3 channels) to regress the light distribution, light intensity and ambient term, respectively. The estimated illumination parameters are fed to the *Neural Projector* for illumination map generation.

5.3.1 Regression Network

The structure of the regression network is shown in Fig. 5.2. The regression network aims to estimate three set of our decomposed illumination parameters including ***light distribution*** P, ***light intensity*** I and ***ambient term*** A, which will be explained as below. For clarity, we take one channel of RGB images as an example in the following description. We first separate light sources from illumination maps since light sources in scenes are most critical in illumination prediction. Following [3], we separate the light source by taking the 5% pixels that have the highest values within the HDR illumination map. The ***Light intensity*** I can then be determined by the summation of all the pixels within the light sources region, and the the ***ambient term*** A is further determined by the averaged pixel value within the remaining region (excluding light sources). We then employ Vogel's method [246] to generate N (N=128 by default in this work) uniformly distributed anchor points on a unit sphere. The value of pixels within the light source region will be assigned to the anchor point with minimum radian distance. The value

of each anchor point is the summation of all assigned pixel values. The value of all anchor points will be normalized by the intensity I to ensure their summation equals one, so that the N anchor points form a standard discrete distribution on a unit sphere as denoted by ***light distribution*** P.

Three branches as shown in Fig. 5.2 are adopted to regress the three sets of parameters respectively. For the light intensity I and ambient term A, a naive L2 loss can be adopted for the regression. But for the light distribution P which are localized on a sphere, a naive L2 loss cannot effectively utilize the spatial information of spherical distribution and the property of standard distribution (the summation of all anchor point values equals one). We take advantage of the subtleties of spherical distribution and propose a novel spherical mover's loss to regress light distribution.

5.3.2 Spherical Mover's Loss

A naive method to predict the discrete spherical distribution is using L2 loss or cross-entropy loss to regress the values of N anchor points, but this naive method often introduces various problems. Firstly, L2 loss only regresses each anchor point separately and cannot effectively evaluate the discrepancy between two sets of anchor points (two distributions). Secondly, both L2 loss and cross-entropy loss cannot effectively utilize the spatial information of the discrete light distribution which is localized on the spherical surface.

Inspired by the Earth Mover's distance which measures the discrepancy between two distributions, we propose a novel spherical mover's loss (SML) to measure the discrepancy between two discrete spherical distributions. To derive the SML, we define two discrete distributions with N points on the sphere as denoted by U and V. Intuitively, SML can be treated as the minimum amount of work required to transform U into V, where the work is measured by multiplying the amount of distribution to be moved and the distance to be moved. Then a transportation plan (or moving plan) matrix T with size of (N, N) can be defined, where each entry T_{ij} in T represents the the amount of probability moved between point U_i and point V_j. Besides, a cost matrix C with size of (N, N) is also defined where each entry C_{ij} in C gives the distance of moving U_i to V_j. Specifically, the distance between a point U_i and another point V_j is measured by their radian distance along the

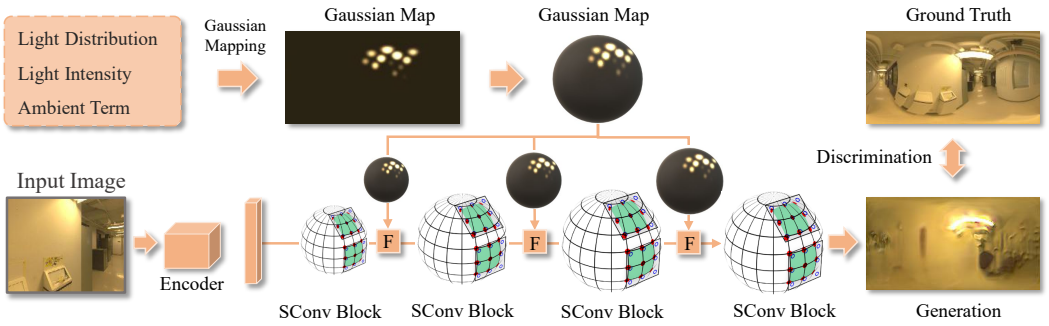


FIGURE 5.3: The structure of neural projector: *SConv Block* denote spherical convolution block, F denotes feature fusion blocks, and *Gaussian Map* is acquired through spherical Gaussian mapping according to the predicted parameters. The *Input Image* is fed to an *Encoder* to produce a feature vector for the ensuing spherical generation. The *Gaussian Map* is fused with the multi-scale spherical generation to synthesize the final illumination map.

unit sphere. As the N anchor points on the sphere surface are pre-defined by the Vogel's method [246], the cost matrix C can be easily pre-computed as a constant matrix in training. With the defined transportation plan matrix T and cost matrix C, SML can be formulated as follows:

$$L_{sml} = \min_{T}(\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}T_{ij}) = \min_{T}\langle C, T\rangle, \quad T\cdot\vec{1} = U,\ T^{\top}\cdot\vec{1} = V \tag{5.1}$$

To solve this problem in a differentiable way, an entropic regularization term $H(T)$ is introduced as defined by $H(T) = -\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}\log T_{ij}$. Then the original problem can be formulated as below:

$$L_{sml} = \min_{T}\langle C, T\rangle - \epsilon H(T), \quad T\cdot\vec{1} = U,\ T^{\top}\cdot\vec{1} = V \tag{5.2}$$

where ϵ is the regularization coefficients which denote the smoothness of the transportation plan matrix T. In our model, ϵ is set to 0.0001 empirically. The regularized form of the problem can be solved efficiently by Sinkhorn iteration [247] in a differentiable way.

Using SML for the regression of spherical distribution has two clear advantages. First, it makes the regression sensitive to the global geometry, thus effectively penalizing predicted activation that is far away from the ground truth distribution. Second, SML is smooth in training which enables stable optimization which is beneficial to the under-constraint problem in illumination prediction.

5.3.3 Neural Projector

With the predicted light distribution, light intensity and ambient term, we propose a Neural Projector to formulate the synthesis of illumination map as a conditional image generation task with paired data as illustrated in Fig. 5.3. To synthesize realistic frequency information in the illumination map, the neural projector is trained in an adversarial manner. The input to the neural projector includes the predicted illumination parameters and the input image. Firstly, we map the parameters into a Gaussian map through spherical Gaussian function [3] as shown below:

$$M = \sum_{i=1}^{N} v_i * exp\frac{d_i * u - 1}{s} + A \tag{5.3}$$

where M is the Gaussian map, N is the number of anchor points, v_i denotes the RGB value of an anchor points which is the product of light distribution on this anchor point and light intensity (namely $v_i = P_i * I$), d_i is the direction of an anchor point (pre-defined by Vogel's method [246]), u is a unit vector giving a direction on the sphere, s is the angular size (selecting 0.0025 empirically), A is the ambient term. Then the constructed Gaussian map will be treated as a guidance (or a condition) for the following generation.

The overall architecture of the generation part is similar to SPADE [27] as shown in Fig. 5.3. Instead of sampling a random vector, we encode the input image as a latent feature vector for the adversarial generation. The illumination map is panoramic image and pixels at different latitudes of a panorama are stretched in different scales. As a result, normal convolution suffers from distortions heavily at different latitudes especially around the polar regions of the panoramic image. Previous work SphereNet [245] builds on regular convolutional filters, which naturally enables the transfer of CNNs between different image representations by adapting the sampling locations of the convolution kernels. We therefore adopt spherical convolution (SConv Block) for the generation of panoramic illumination map, effectively reversing distortions and wraps the filters around the sphere. The Gaussian map is then fused with the feature of SConv Block in multiple scales through the spatially adaptive normalization as described in [27].

The neural projector employs several losses to drive the generation of high-quality illumination maps. We denote the input Gaussian map as x, the ground-truth illumination map as y, and the generated illumination map as x'. To stabilize

the training, we introduce a feature matching loss $\mathcal{L}_{feat}$ to match the intermediate features of discriminator between the generated illumination map and ground truth:

$$\mathcal{L}_{feat} = \sum_{l} \lambda_l ||D_l(x, x') - D_l(x, y)||_1 \tag{5.4}$$

where D_l represents the activation of layer l in the discriminator and λ_l denotes the balanced coefficients. To obtain a similar illumination distribution instead of excessively emphasizing the absolute intensity, a cosine similarity is computed between the generated illumination map and ground truth as below:

$$\mathcal{L}_{cos} = (1 - Cos(x', y)) * \lambda_{cos} \tag{5.5}$$

where λ_{cos} is the weight of this term. The discriminator adopts the same architecture with Patch-GAN [25], thus obtaining the adversarial loss of discriminator D and generator G as denoted by $\mathcal{L}_D$ and $\mathcal{L}_G$. Then the neural projector is optimized following the objective as below:

$$\mathcal{L} = \min_G \max_D (\mathcal{L}_{feat} + \mathcal{L}_{cos} + \mathcal{L}_G + \mathcal{L}_D) \tag{5.6}$$

As the regression network and neural projector are all differentiable, the whole framework can be optimized end-to-end.

5.3.4 Experiments

Dataset and Experimental Setting: We evaluate EMLight with the Laval Indoor HDR Dataset [61] that consists of 2,100 HDR panoramas taken in a variety of indoor environments. Similar to Gardner et al. [61], we crop eight images with limited field of views from each panorama which produces 19,556 training pairs as used in our experiments. For each of the 19,556 images, the same image warping operation in Gardner et al. [61] is applied. In the experiments, we randomly select 200 images as the testing set and the rest for training.

Consistent with Gardner et al. [3] and Garon et al. [2], DenseNet121 is used as the backbone in regression network.

Evaluation Method and Metric: Similar to the evaluation setting in DeepLight [64], our scenes for evaluations include three spheres with different materials:



FIGURE 5.4: The scenes used in evaluations consist of three spheres with different materials including diffuse gray, matte silver and mirror silver.

TABLE 5.1: Comparison of EMLight with several state-of-the-art lighting estimation methods: The evaluation metrics include the widely used RMSE, siRMSE, Angular Error (AE) and AMT. D, S, M denote a diffuse, a matte silver and a mirror material of the rendered objects, respectively.

Metrics	[61]			[3]			[63]			[2]			**EMLight**		
	D	S	M	D	S	M	D	S	M	D	S	M	D	S	M
RMSE	0.146	0.173	0.202	0.084	0.112	0.147	0.203	0.218	0.257	0.181	0.207	0.249	**0.062**	**0.071**	**0.089**
siRMSE	0.142	0.151	0.174	0.073	0.093	0.119	0.193	0.212	0.243	0.177	0.196	0.221	**0.043**	**0.054**	**0.078**
AE	8.12°	8.37°	8.81°	6.82°	7.15°	7.22°	9.37°	9.51°	9.81°	9.12°	9.32°	9.49°	**6.43°**	**6.61°**	**6.95°**
AMT	28.0%	23.0%	20.5%	33.5%	28.0%	24.5%	25.0%	21.5%	17.5%	27.0%	22.5%	19.0%	**40.0%**	**35.5%**	**31.0%**

gray diffuse, matte silver and mirror as illustrated in Fig. 5.4. The performance is evaluated by comparing the scene images rendered (by Blender [4]) with predicted illumination maps and ground truth. The evaluation metrics include Root mean square error (RMSE) and scale-invariant RMSE (si-RMSE) that focus on the estimated light intensity and light directions (or shadings), respectively. Both metrics have been widely adopted in the evaluation of illumination prediction. In addition, we also adopt the per-pixel linear RGB angular error [64] and Amazon Mechanical Turk (AMT) which performs crowd-sourcing user study for subjective assessment of empirical realism of rendered images. In the experiments, each compared model predicts 200 illumination maps on the testing set for quantitative evaluation. For qualitative evaluation, we design 25 scenes for 3D insertion and render them with the predicted illumination maps.

Quantitative Evaluation: We compare EMLight with several state-of-the-art methods that estimate illumination maps directly [61] or estimate representative illumination parameters [2, 3, 63]. For each compared method, we render 200 images of the testing scene (three spheres with diffuse, matte silver, mirror silver materials) by using the illumination maps predicted from the testing set. Table

5.1 shows experimental results, where D, S and M denote diffuse, matte silver and mirror material of the objects to be rendered, respectively. AMT user study is conducted by showing two images rendered by the ground truth and each compared methods in Table 5.1 to 20 users who will pick more realistic image. The score is the percentage of rendered images (200 images in total) that is deemed as more realistic than the ground-truth rendering.

We can observe that EMLight outperforms all compared methods under different evaluation metrics and materials consistently, largely attributed to the accurate generation of illumination under the guidance of the Gaussian map. Gardner et al. [61] predict illumination maps directly but the direct generation without any guidance tends to over-fit training data and lead to sub-optimal generalization due to the unconstrained nature of illumination estimation from a single image. Gardner et al. [3] regress spherical Gaussian parameters of light sources which tends to lose useful frequency information and generate inaccurate shading and shadow as measured by si-RMSE. Li et al. [63] adopt spherical Gaussian functions to reconstruct the full illumination map in the spatial domain which often fails to recover high-frequency illumination. Garon et al. [2] recover lighting by regressing spherical harmonic coefficients while it struggles to regress accurate light directions and recover high-frequency information. Though a masked L2 loss is employed in Garon et al. [2] for preserving high-frequency information, it does not solve the problem essentially as illustrated in Fig. 5.5. As a comparison, EMLight estimates illumination parameters accurately by regressing light distribution under a spherical mover's loss. Under the guidance of estimated parameters, the neural projector generates accurate and high-fidelity illumination maps with realistic frequency information through adversarial training.

Qualitative Evaluation: We visualize our predicted Gaussian maps, generated illumination maps, and the corresponding intensity maps in Fig. 5.6. As Fig. 5.6 shows, our regression network predicts light distribution accurately as shown in *Gaussian Map*. The neural projector generates accurate and realistic HDR illumination maps as shown in *Generation*. To further verify the quality of generated HDR illumination, we visualize the intensity maps of the illumination maps.

We compare EMLight with four state-of-the-art light estimation methods qualitatively. Fig. 5.5 shows rendered images with the predicted illumination maps

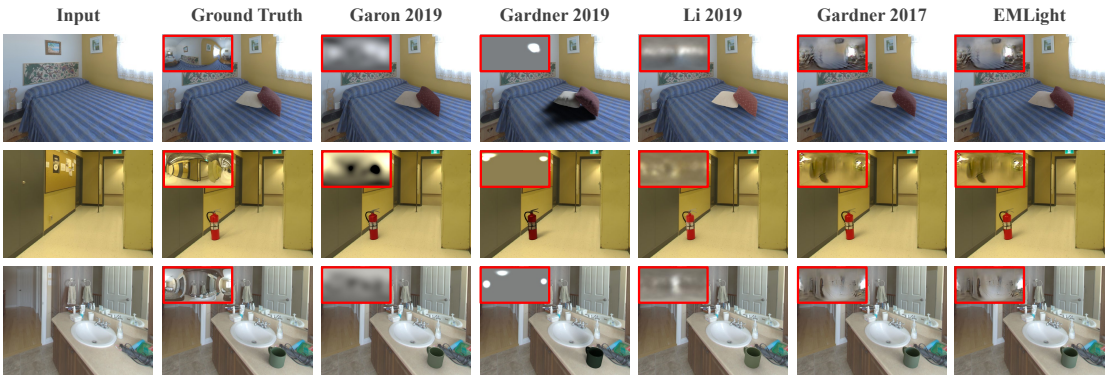


FIGURE 5.5: Visual comparison of EMLight with state-of-the-art lighting estimation methods: With the illumination maps predicted by different methods (at top-left corner of each rendered image), the rendered virtual objects demonstrate different light intensity, color, shadow and shade.

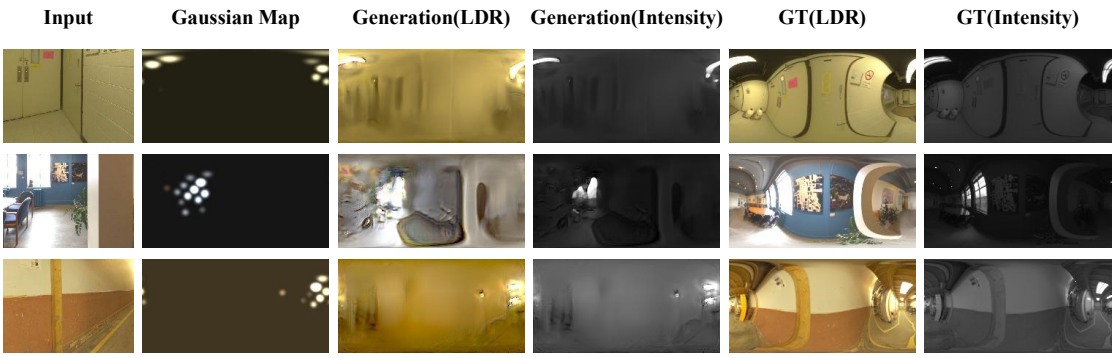


FIGURE 5.6: Illustration of EMLight illumination estimation: For the input images in column 1, columns 2 shows the constructed Gaussian maps based on the regressed illumination parameters and columns 3 and 4 show the generated illumination map under the guidance of Gaussian map and the corresponding intensity map, respectively. Columns 5 and 6 show the ground truth of the illumination maps and the corresponding intensity maps, respectively.

(highlighted by red boxes). We can observe that EMLight predicts realistic illumination maps with plausible light sources and produces realistic rendering with clear and accurate shade and shadows. As a comparison, direct generation [61] struggles to identify the direction of light sources as there is no guidance for the generation. Illumination maps by Gardner et al. [3] are over-simplified with a limited number of light sources, and the simplification loses accurate frequency information which results in unrealistic shadow and shading in rendering. Garon et al. [2] and Li et al. [63] regress illumination parameters but are often constrained by the order of representative functions (spherical harmonic and spherical Gaussian). As a result, they predict illumination of low frequency and produce renderings with very weak shade and shadow in rendering as illustrated in Fig. 5.5.

TABLE 5.2: Ablation study of the proposed EMLight: SG and SD denote the spherical Gaussian representation and our spherical distribution representation of the illumination map. L2 and SML denote using the L2 and spherical mover's loss to regress the representation parameters. NP denotes the proposed neural projector.

Models	**RMSE**			**si-RMSE**			**Angular Error**			**AMT**		
	D	S	M	D	S	M	D	S	M	D	S	M
SG+L2	0.204	0.213	0.238	0.188	0.203	0.229	9.18°	9.42°	9.73°	26.0%	22.5%	18.0%
SD+L2	0.133	0.161	0.178	0.117	0.132	0.161	7.60°	7.88°	8.12°	30.5%	25.5%	22.0%
SD+SML	0.080	0.103	0.117	0.072	0.087	0.106	6.78°	6.98°	7.12°	34.0%	31.5%	26.0%
SD+SML+NP	**0.062**	**0.071**	**0.089**	**0.043**	**0.054**	**0.078**	**6.43°**	**6.61°**	**6.95°**	**40.0%**	**35.5%**	**31.0%**

Ablation Study: We develop several EMLight variants as listed in Table 5.2 to evaluate the effectiveness of proposed methods. The variants include 1) the baseline *SG+L2* that regresses spherical Gaussian representative parameters with L2 loss as in [61]; 2) the *SD+L2* that regresses the spherical distribution of illumination with L2 loss; 3) the (*SD+SML*) that regresses the spherical distribution of illumination with SML; and 4) the standard model *SD+SML+NP*. Similar to the setting in *Quantitative Evaluation*, we apply all variant models to render 200 images of the testing scene. As Table 5.2 shows, (*SD+L2*) outperforms *SG+L2* clearly, demonstrating the superiority of the spherical distribution representation of illumination. *SD+SML* also produces better estimation than *SD+L2*, validating the effectiveness of SML. *SD+SML+NP* achieves the best estimation, demonstrating that the neural projector promotes the performance of illumination prediction significantly.

We also benchmark spherical mover's loss (SML) with the widely adopted Cross-Entropy Loss for distribution regression, compare the spherical convolution with normal convolution, and study how anchor points affect the light estimation as shown in Table 5.3. We followed the experimental setting as in Table 5.2 and measure the averaged RMSE and si-RMSE on three materials. As Table 5.3 show, SML outperforms cross-entropy loss clearly as SML captures spatial information of spherical distributions effectively. In addition, spherical convolution performs better than normal convolution consistently in panoramic image generation. Further, the prediction performance drops slightly when 64 instead of 128 anchor points are used and increasing anchor points to 196 doesn't improve the performance obviously. We conjecture that the larger number of parameters with 196 anchor points affects the regression accuracy negatively.

TABLE 5.3: Ablation studies over anchor points, loss functions and convolution types: EMLight* denotes the standard EMLight with 128 anchor points, spherical mover's loss (SML), and spherical convolution. We create four EMLight variants by setting the number of anchor points to 64 and 196, replacing SML with cross-entropy loss, and replacing spherical convolution with normal convolution.

Models	RMSE	si-RMSE
Anchor=64	0.091	0.075
Anchor=196	0.076	**0.055**
Cross-Entropy Loss	0.102	0.082
Normal Convolution	0.086	0.071
EMLight*	**0.074**	0.058

5.4 Sparse Needlets for Lighting Estimation

Lighting estimation has been tackled by regressing the parameters of various lighting representations in either frequency domain [2, 59]) or spatial domain [3, 61, 62, 65]. However, lighting estimation in frequency domain normally represents illumination with Spherical Harmonics (SH) which lack spatial localization capabilities. Thus, it tends to capture global lighting instead of the exact spatial locations of the light sources which often leads to weak shading and shadow effects as illustrated in Garon et al. [2] of Fig. 5.7. Lighting estimation in spatial domain has been addressed by direct generation of the illumination maps or indirect reconstruction through spherical Gaussian function. However, direct generation of illumination maps often leads to worse generalization as lighting estimation is an under-constrained problem by itself, and spherical Gaussian often involves a complicated training process as described in [3]. Both types of approaches in spatial domain do not explicitly consider lighting frequency, and thus lead to inaccurate relighting performance as illustrated in Gardner et al. [3] of Fig. 5.7. The high frequency information also tends to be blurred due to the use of naive L2 loss in the training. Additionally, existing evaluation metrics in lighting estimation usually assess the objects rendered with the predicted illumination maps, which is time-consuming and sensitive to the test setting.

We propose NeedleLight [38] [4], a new model that introduces needlet for accurate and robust lighting estimation from a single image. As a new generation of spherical wavelets, needlet enjoys good localization properties in both frequency and

[4]The work is under submission



Garon et al. [2] Gardner et al. [3] Ours Ground Truth

FIGURE 5.7: The proposed NeedleLight estimates a parametric lighting representation from a single scene image which is critical to many tasks such as virtual object insertion. Unlike previous methods that predict lighting in either frequency domain [2] (losing spatial localization) or spatial domain [3] (losing frequency information) only, we introduce a novel needlets basis which is capable of representing and estimating lighting accurately in both frequency and spatial domains.

spatial domain which makes it ideal to be the basis for illumination representation. Moreover, to remove the redundant parameters in needlet coefficients which will disturb the regression of principle light sources, we design an optimal thresholding function to achieve sparse needlets which improve the lighting estimation greatly. Leveraging the spatial localization of needlet coefficients,

Unlike spherical harmonic coefficients, needlet coefficients are spatially localized over a unit sphere. To utilize the spatial information in regression, we propose a Spherical Transport Loss (STL) based on optimal transport theory. STL is able to capture spatial information via a cost matrix and estimate the needlet coefficients more accurately than a naive L2 loss. Besides, STL employs auxiliary point strategy to preserve high frequency information and greatly reduce the dimension of required parameters. Based on STL, we design a new metric for the evaluation of lighting estimation by measuring the discrepancy between illumination maps. The new metric highly simplifies the evaluation procedure and provides concise yet effective evaluation with regard to the lighting color, intensity and position.

The contribution of this work can be summarized in three aspects. First, we introduce a novel needlet basis for illumination representation which allows to regress the parameters in both frequency and spatial domains simultaneously. Second, we develop an optimal thresholding function to achieve sparse needlets which effectively removes the redundant needlet coefficients and improves the lighting estimation. Third, we design a novel Spherical Transport Loss (STL) that effectively utilizes

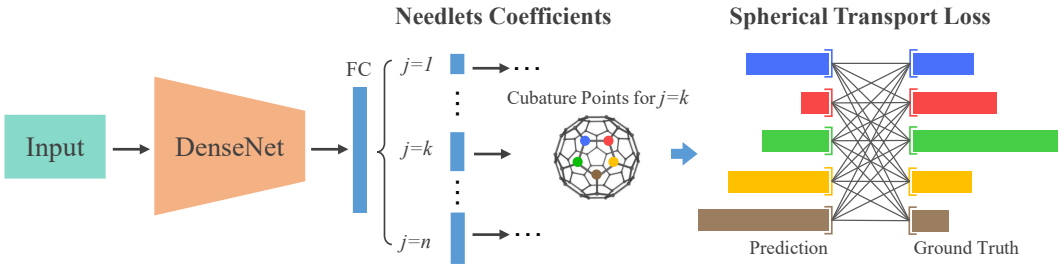


FIGURE 5.8: The architecture of the proposed NeedleLight: We introduce sparse needlets to represent lighting. The frequency bands in *Needlet Coefficients* are denoted as $j = 1, \cdots, n$. The needlet coefficients in each frequency band are spatially localized on a set of *Cubature Points* on the unit sphere (illustration for $j = k$ only). For stable lighting regression, we design a spherical transport loss to capture the divergence between the predicted needlet coefficients (*Prediction*) and the ground-truth coefficients (*Ground Truth*). The values of needlet coefficients are illustrated by color bars (only draw 5 for illustration).

the spatial information of needlet coefficients in regression. With STL, we also design a new evaluation metric that is more concise and effective than existing evaluation metrics.

The proposed NeedleLight estimates illumination by regressing needlet coefficients from a single image as illustrated in Fig. 5.8. A novel spherical transport loss is designed to achieve stable and effective regression of needlet coefficients. The following subsections describe how needlets and spherical transport loss work together for accurate and robust lighting estimation.

5.4.1 Needlets-based Lighting Representation

Needlets [240, 241] are a new type of spherical wavelets that has been successfully used in microwave signal analysis. They can be localized at a finite number of frequencies, and decay quasi-exponentially fast away from the global maximum. Thus, they enjoy good localization properties in both frequency and spatial domains. As described in [241], a signal $I(x)$ (e.g. lighting signals of interest in this research) at a given frequency $j \in \mathbb{N}$ can be represented by the spherical needlets

basis $\psi_{jk}(x)$ and the needlet coefficients β_{jk} as follows:

$$\begin{aligned}\psi_{jk}(x) &= \sqrt{\lambda_{jk}} \sum_{l=\lceil B^{j-1} \rceil}^{\lfloor B^{j+1} \rfloor} b(\frac{l}{B^j}) \sum_{m=-l}^{l} Y_{lm}(\xi_{jk}) \overline{Y}_{lm}(x) \\ \beta_{jk} &= \sqrt{\lambda_{jk}} \sum_{l=0}^{\infty} b(\frac{l}{B^j}) \sum_{m=-l}^{l} a_{lm} Y_{lm}(\xi_{jk})\end{aligned} \tag{5.7}$$

where $x \in \mathbb{S}^2$, ξ_{jk} and λ_{jk} are pre-defined cubature points that spread over the unit sphere as shown in Fig. 5.9, and the associated cubature weights, respectively, $b(\cdot)$ is a window function, B is a free parameter larger than 1, Y_{lm} is spherical harmonic function with degree l and order m, a_{lm} is the corresponding spherical harmonic coefficients, $\overline{Y}_{lm}$ is the complex conjugation of Y_{lm}. ξ_{jk} represents the spatial location of needlet basis ψ_{jk}, thus the needlet coefficients are spatially localized on the unit sphere. The signal $I(x)$ can be reconstructed via $I(x) = \sum_{j,k} \beta_{jk} \psi_{jk}(x)$.

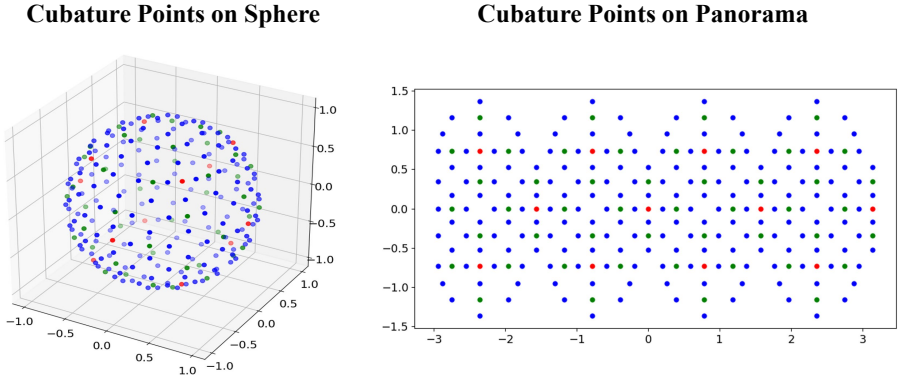


FIGURE 5.9: The visualization of cubature points (j_{max}=3) on sphere and panorama. The red, green and blue points denote the cubature points of frequency bands from low to high frequency, namely j=1, j=2 and j=3 respectively.

Compared with SH, needlets have compact supports for the localization in spatial domain. As a result, they can easily and parsimoniously represent signals over the unit sphere that exhibits local sharp peaks or valleys, which are commonly presented in HDR illumination maps. Consequently, needlets serve as a more suitable basis for the representation of illumination maps.

5.4.2 Sparse Needlets

A signal is said to be sparse if it can be reconstructed with only a small amount of basis functions. An illumination map consists of several dominant light sources

with high radiance energy and an ambient residual. The light sources are obviously sparse under the needlet basis while the ambient component is not. For the lighting estimation, the reconstruction of dominant light sources is more significant than the reconstruction of the ambient component especially in high frequency sections. Besides, there are dozens of needlet coefficients in high frequency section (252 coefficients for j_{max}=3), most of which are redundant parameters of ambient. Those redundant parameters will severely disturb the regression of principal light sources and lead to diffuse illumination (or low frequency illumination) which is undesired for relighting. Thus, we deduce an optimal sparse function for needlets to separate the principle coefficients of light sources from the redundant ambient component.

We derive the sparse function from a Bayesian perspective and form the problem as a maximum posterior estimator. Normally, we assume that the needlet coefficients of light sources s follows Laplace distribution prior as Laplace prior is well adapted to model sparse signals [248], which is also proved in [249]. The needlet coefficients of ambient can be ideally treated as a Gaussian distribution [250]. The needlet coefficients β of the illumination map can thus be modeled as:

$$\beta = \underbrace{s}_{light\ sources} + \underbrace{\phi}_{ambient} + \underbrace{\eta}_{noise} \tag{5.8}$$

where s denotes the needlet coefficients of sparse light sources which follow Laplace distribution, ϕ denotes the needlet coefficients of ambient which is a Gaussian distribution, and η denotes noises that follow a Gaussian distribution. According to Vansynge et al. [251], the Bayesian formulation of the problem can be written as:

$$P(s|\beta) \propto P(s)\mathcal{L}(\beta|s), \quad \mathcal{L}(\beta|s,\phi) = N(\beta; \phi + s, M_\eta) \tag{5.9}$$

where M_η denotes the covariance matrices of the noise. As ϕ follows Gaussian distribution, we can derive:

$$\begin{aligned}\mathcal{L}(\beta|s) = \int ... \int \mathcal{L}(\beta|s,\phi)P(\phi)d\phi \propto \exp\left[-\frac{1}{2}\beta^T M_\eta^{-1}\beta + s^T M_\eta^{-1}\beta - \frac{1}{2}s^T M_\eta^{-1}s\right] \\ \cdot \exp\left[\frac{1}{2}(M_\eta^{-1}\beta - M_\eta^{-1}s)^T(M_\eta^{-1} + M_\phi^{-1})(M_\eta^{-1}\beta - M_\eta^{-1}s)\right]\end{aligned} \tag{5.10}$$

where M_ϕ denotes the covariance matrices of the Gaussian distribution ϕ. As $P(s)$ follows the Laplace distribution namely $P(s) \propto \exp[-\lambda||s||]$, the maximum posterior estimator is obtained by maximizing $P(s|\beta) = \mathcal{L}(\beta|s) * P(s)$. We take

the partial derivative with respect to s:

$$\begin{aligned}\partial_s(-log(P(s|\beta))) = &-(M_\eta + M_\eta M_\phi^{-1} M_\eta)^{-1} s + M_\eta^{-1} s \\ &+ (M_\eta + M_\eta M_\phi^{-1} M_\eta)^{-1} \beta - M_\eta^{-1} \beta + \lambda \partial_s ||s||\end{aligned} \tag{5.11}$$

By making $\partial_s(-log(P(s|\beta))) = 0$, the following solution (or sparse function) can thus be derived:

$$s = \beta - \left[M_\eta^{-1} - (M_\eta + M_\eta M_\phi^{-1} M_\eta)^{-1}\right]^{-1} \lambda \partial_s ||s|| \tag{5.12}$$

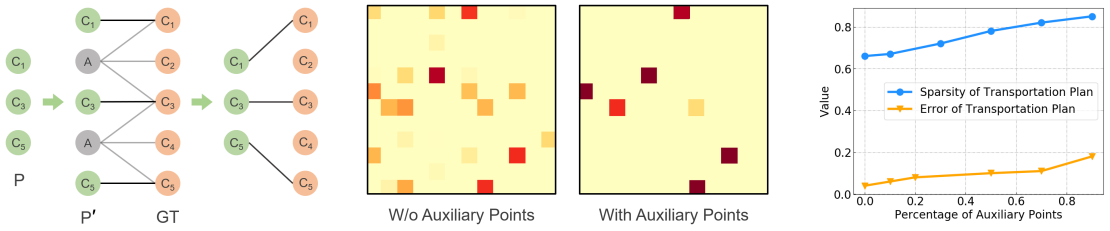
which is a soft thresholding operator with threshold $(M_\eta^{-1} - (M_\eta + M_\eta M_\phi^{-1} M_\eta)^{-1})\lambda$.

We apply the sparse function largely to the high frequency needlet coefficients so as to shrink the redundant coefficients of the ambient component. The sparse needlets provide sparse representation of illumination maps, which is desirable for the regression of light sources by using the proposed spherical transport loss to be described in the following subsection.

5.4.3 Spherical Transport Loss

Different from spherical harmonics, needlet coefficients are spatially localized in a unit sphere. A simple MSE loss cannot utilize the spatial information. Besides, as an under-constrained task, the optimization in lighting estimation is severely challenging since it aims to recover the environment lighting from all directions based on a single scene image with a limited field of view. Thus the training of lighting estimation models struggles with the regression of high frequency information.

In this work, we propose a novel Spherical Transport Loss (STL) to achieve the stable and effective lighting regression. Because the needlet coefficients are spatially localized on the unit sphere as specified by cubature points, we treat the regression of needlet coefficients as an Unbalanced Optimal Transport (UOT) [252, 253] problem on the unit sphere. Intuitively, UOT computes the cost for transporting a measure distributed on a space to another measure of possibly different total masses. As the needlet coefficients may contains negative values, we take the natural logarithm of them before deriving the spherical transport loss. Then we can define two sets of needlet coefficients represented by two positive vectors $a = (a_1, \cdot, a_n) \in \mathbb{R}_+^n$ and $b = (b_1, \cdots, b_n) \in \mathbb{R}_+^n$ and their spatial layouts are specified by cubature points on the unit sphere. A distance matrix C and a transportation plan matrix P can thus be derived, where each entry C_{ij} in C gives the cost of moving point a_i to point b_j which can be defined by radian distance between points on the unit sphere,



(a) (b) (c)

FIGURE 5.10: (a) The process to achieve sparse transportation plan through auxiliary points. P and GT are the predicted needlet coefficients and the ground truth, respectively. 'A' denotes the auxiliary point. P′ denotes the predicted coefficients extended with auxiliary points. The connection between P and GT signifies the resulting transportation plan. (b) Transportation plans for the needlet coefficients (j=1) w/o and w/ auxiliary points, respectively. (c) The sparsity and error of transportation plan matrix with different percentage of auxiliary points.

and P_{ij} in P represents the probability of assigning a point a_i to a point b_j. A regularized UOT problem can thus be defined as follows:

$$\min_{P} \left[\langle C, P \rangle + \tau KL(P \cdot \vec{1} || a) + \tau KL(\vec{1} \cdot P || b) - \eta H(P) \right] \tag{5.13}$$

where τ and η are regularization parameters, KL is the KerKullback-Leibler Divergence, $H(P)$ is the entropic regularization for efficient approximation of original UOT problem [253]. The regularized UOT can be solved by Sinkhorn iteration [247].

To construct clear and sharp instead of diffuse (i.e., low frequency) light sources, we expect the transportation plan to be sparse. Otherwise the diffuse light sources will result in weak shading and shadow effects as presented in Garon et al. [2] of Fig. 5.12. Besides, the dimension of output layer increases quickly when the frequency of needlets becomes large (252 needlet coefficients when $j_{max} = 3$). We propose an auxiliary point strategy to achieve sparse transportation plan and reduce the dimension of output layer. Auxiliary points are assigned with small value (we select $\frac{1}{(number\ of\ coefficients)}$) and 0 cost for transport, which can be used for absorbing unused probability mass in cases of partial transport. As shown in Fig. 5.10(a), we only predict partial needlet coefficients (P), and then use the auxiliary points to replace other needlet coefficients to obtain a new set of coefficients (P′). After obtaining the optimal transport between P′ and the ground truth GT, we can extract a sparse transportation plan by removing the connection with the auxiliary points. The two samples in Fig. 5.10(b) show the transportation plan matrix for

the needlet coefficients ($j = 1$) with and without auxiliary points, respectively. Fig. 5.10(c) shows the sparsity and error of the transportation plan matrix ($n \times n$) with different percentage of auxiliary points (The evaluation metrics for the sparsity and error of transportation plan are described in [254]). When more auxiliary points are used, the sparsity of the transportation matrix increases but the error also increases as shown in Fig. 5.10 (c). As a trade-off, we select 66% of coefficients as auxiliary points.

Using spherical transport loss for needlet coefficients regression has two clear advantages. First, it makes the regression sensitive to the global geometry, thus effectively penalizing predicted activation that is far away from the ground truth distribution. Second, it can preserve the high frequency information during training with the proposed auxiliary points.

Evaluation Metric: Lighting estimation has been widely evaluated by using root mean square error (RMSE) and scale-invariant RMSE (si-RMSE) that measure the standard deviation of residuals of the rendered images. RMSE mainly evaluates the estimated lighting intensity, and si-RMSE focuses more on the evaluation of lighting positions. In addition, lighting estimation has also been evaluated by using Amazon Mechanical Turk (AMT) that performs crowd-sourcing user study for subjective assessment of empirical realism of rendered images.

Existing metrics evaluate estimated lighting largely by applying them to the rendered objects. Thus the performance of an estimation model is highly affected by test settings such as materials and 3D shape of the rendered objects. Based on the proposed spherical transport loss, we design a spherical transport distance (STD) metric that directly evaluates the optimal transport distance between the predicted illumination map and the ground-truth map on the unit sphere. The only difference from STL is that STD discards the auxiliary point strategy. The proposed STD highly simplifies the evaluation procedure and provides concise yet effective evaluations regarding to the lighting color, lighting intensity, and lighting position jointly.

5.4.4 Experiments

Dataset and Implementation: We evaluate NeedleLight by using the Laval Indoor HDR Dataset [61] that consists of 2,100 HDR panoramas taken in a variety of indoor environments. Similar to [61], we extract eight limited field of view crops from each panorama which produces 19,556 images as used in our experiments. The image warping operation as described in [61] is applied to the panoramas. We apply the proposed sparse needlets with $j_{max} = 3$ to extract needlet coefficients as the ground truth for training. Similar to [2, 3], DenseNet121 is used as the backbone network and produces a 4096-dimensional latent vector which is further forwarded to a fully-connected layer with 1024 units. Three separate output layers are added to regress needlet coefficients in frequency bands $j = 1, 2, 3$. In the experiment, we randomly select 200 images as testing set and the rest for training set. All the objects are rendered with Blender [4].

The proposed NeedleLight is implemented by the PyTorch framework. The Adam is adopted as optimizer which employs a learning rate decay mechanism (initial learning rate is 0.001). The network is trained in 100 epochs with a batch size of 64. In addition, the network training is performed on two NVIDIA Tesla P100 GPUs with 16GB memory.



FIGURE 5.11: The scene for quantitative evaluations consists of three spheres with diffuse gray, matte silver and mirror silver materials.

Quantitative Evaluation: We compare NeedleLight with a number of state-of-the-art lighting estimation methods including Garon et al. [2] that estimates lighting in frequency domain and Gardner et al. [3, 61] and Li et al. [63] that estimate lighting in spatial domain. To perform quantitative evaluations, we create three spheres with gray diffuse, matte silver and mirror silver materials for evaluation as shown in Fig. 5.11, which is consistent with the evaluation setting in [64].

Then we render 300 images of objects (100 images for each material) by using the illumination maps that are predicted from testing set by each compared method. Table 5.4 shows experimental results by using 4 evaluation metrics as described in *Evaluation Metrics*, where D, S and M denote a diffuse, a matte silver and a mirror material of the objects to be rendered, respectively. The AMT user study is conducted by showing two images rendered by the ground truth and one of the methods in Table 5.4 to 20 users who will pick the more realistic image. The score is the percentage of images rendered by the method that is deemed as more realistic than the ground-truth rendering.

TABLE 5.4: Comparison of NeedleLight with several state-of-the-art lighting estimation methods: The evaluation metrics include the widely used RMSE, si-RMSE, AMT, and our proposed STD. D, S, M denote a diffuse, a matte silver and a mirror material of the rendered objects, respectively.

Metrics	**[61]**			**[3]**			**[63]**			**[2]**			**NeedleLight**		
	D	S	M	D	S	M	D	S	M	D	S	M	D	S	M
RMSE	0.13	0.16	0.18	**0.06**	0.10	0.15	0.21	0.23	0.26	0.18	0.20	0.24	0.07	**0.07**	**0.09**
si-RMSE	0.15	0.15	0.17	0.07	0.09	0.12	0.19	0.21	0.23	0.21	0.24	0.26	**0.05**	**0.06**	**0.08**
AMT	28%	23%	21%	34%	33%	30%	28%	27%	23%	29%	26%	24%	**41%**	**39%**	**36%**
STD	6.84			5.52			7.01			7.14			**4.21**		

We can observe that NeedleLight outperforms other methods in most cases under different evaluation metrics and materials as it allows regression in frequency and spatial domain jointly. The only exception is for diffuse material by [3] while evaluated using RMSE, largely because the parameterization in Gardner et al. [3] simplifies the scene illumination in spatial domain to achieve accurate prediction of light intensity while diffuse material is largely affected by light intensity and RMSE is most sensitive to light intensity. Gardner et al. [61] predicts illumination maps directly by a two-stage training strategy. As an under-constrained problem, the direct generation methods like Gardner et al. [61] tend to over-fit the training set and present worse generalization performance. Besides, both Gardner et al. [61] and Gardner et al. [3] estimate lighting in spatial domain which cannot recover frequency information and tends to generate inaccurate shading and shadow that are largely measured by si-RMSE. Garon et al. [2] recovers lighting in frequency domain by regressing the SH coefficients, which tend to capture global instead of localized lighting. Thus Garon et al. [2] struggles to regress accurate lighting position and recover high frequency information. Li et al. [63] adopts spherical

Gaussian functions to reconstruct illumination in spatial domain, thus it cannot recover accurate illumination frequency. Besides, it uses a masked L2 loss to preserve high frequency information though it cannot solve the missing of high frequency information essentially as illustrated in Fig. 5.12. Instead, our proposed spherical transport loss with auxiliary points improves the regression of high frequency information significantly.

In addition, we can observe that the performance of the state-of-the-art methods is not consistent under different evaluation metrics. For example, Li et al. [63] outperforms Garon et al. [2] in si-RMSE but the situation becomes the other way around in RMSE. The divergence of different metrics makes it hard to provide consistent evaluations. The proposed spherical transport distance (STD) instead provides relatively consistent and comprehensive evaluations in light intensity, color, position, etc. as shown in Table 5.4.



FIGURE 5.12: Visual comparison of NeedleLight with state-of-the-art lighting estimation methods: With the illumination maps predicted by different methods (at bottom-right corner of each rendered image), the rendered virtual objects demonstrate different lighting intensity, color, shadow and shade.

Qualitative Evaluation: We also compare NeedleLight with four state-of-the-art lighting estimation methods qualitatively. we design 25 3D scenes with objects for insertion and render them with the predicted illumination maps. Fig. 5.12 shows several rendered images and the predicted illumination maps. We can observe that NeedleLight predicts realistic illumination maps with plausible light sources, thus producing realistic rendering with clear and accurate shade and shadows that are very close to the ground truth. As a comparison, the illumination maps predicted by Gardner et al. [61] and Gardner et al. [3] tend to present clearly higher illumination frequency than the ground truth, largely because they recover illumination



FIGURE 5.13: Object relighting on a variety of photos from the Internet. In all cases, light estimation is performed completely automatically by our model. The predicted illumination maps are utilized to relight the virtual objects with Blender [4].

in spatial domain without explicitly taking frequency information into account. Garon et al. [2] and Li et al. [63] only predict illumination of low frequency, thus the produced renderings present very weak shade and shadow effects as illustrated in Fig. 5.12.

Besides the testing set, we also validate the proposed method on natural images collected from the Internet as shown in Fig. 5.13. The proposed method achieves accurate estimation of scene illumination, thus the 3D objects can be embedded into the images with real shading and shadow effects.

TABLE 5.5: Ablation study of the proposed NeedleLight: *SH* denotes using spherical harmonics for representation; *SN* denotes using spherical needlets for representation; *ST* denotes applying the derived soft thresholding function to needlet coefficients; *L2* and *STL* denotes regressing coefficients with L2 loss and spherical transport loss; *SN+ST+L2+STL* denotes standard NeedleLight with sparse needlets which are regressed with both L2 and spherical transport loss.

	RMSE			**si-RMSE**			**AMT**			**STD**
Models	D	S	M	D	S	M	D	S	M	-
SH+L2	0.19	0.20	0.25	0.22	0.24	0.28	23%	21%	18%	7.14
SN + L2	0.14	0.14	0.16	0.11	0.14	0.15	33%	31%	28%	4.93
SN+ST+L2	0.11	0.12	0.14	0.09	0.09	0.13	34%	32%	30%	4.74
SN+ST+STL	0.10	0.11	0.13	0.07	0.10	0.12	37%	34%	32%	4.51
SN+ST+L2+STL	**0.07**	**0.07**	**0.09**	**0.05**	**0.06**	**0.08**	**41%**	**39%**	**36%**	**4.21**

Ablation Study: We further evaluate NeedleLight by developing four NeedleLight variants as listed in Table 5.5, including a baseline model which regress spherical harmonic coefficients with L2 loss (*SH+L2*), regressing original needlet coefficients with L2 loss (*SN+L2*), regressing needlet coefficients after applying soft thresholding (namely sparse needlets) with L2 loss (*SN+ST+L2*), regressing sparse needlets coefficients with spherical transport loss (*SN+ST+STL*). The standard NeedleLight regresses spares needlets with both L2 and spherical transport loss (*SN+ST+L2+STL*). Similar to the setting in Quantitative Evaluation, we apply the five variants and the standard NeedleLight to render 300 images with 15 objects of different materials. As Table 5.5 shows, using needlets for illumination representation (SN+L2) helps to achieve better lighting estimation compared with spherical harmonics function (SH+L2). In addition, the performance of *SN+ST+L2* is improved clearly as sparse needlet coefficients helps the regression of light sources significantly by the zero setting of redundant coefficients. Additionally, the standard *SN+ST+L2+STL* outperforms *SN+ST+L2* and *SN+ST+STL*, demonstrating the L2 and the proposed STL are complementary for the regression of needlet coefficients.

We also study effect of thresholding function and different orders of needlets and compare spherical needlets (SN) with other basis functions for illumination representation such as spherical harmonics (SH), spherical Gaussian (SG) and Haar [255] as shown in Table 5.6. We followed the experimental setting as in Table 5.5 and measure the proposed STD as the evaluation metric. The number of coefficients of SH, SG and Haar are set to be consistent with spherical needlets with 3

TABLE 5.6: Ablation studies over different bases including spherical harmonics (SH), spherical Gaussian (SG), Haar, spherical needlets (SN). HT and ST denote apply hard thresholding and the derived soft thresholding functions to the spherical needlet coefficients. j_{max} denotes the order of needlets for representation. SN(j_{max}=3)+ST* is the standard setting of NeedleLight.

Models	**L2**	**L2 + STL**
SH	7.14	7.13
SG	6.01	5.74
Haar	6.07	5.83
SN (j_{max}=1)	7.52	7.18
SN (j_{max}=2)	6.71	6.20
SN (j_{max}=4)	5.37	4.96
SN(j_{max}=3)	5.32	4.93
SN(j_{max}=3)+**HT**	4.92	4.51
SN(j_{max}=3)+**ST***	**4.74**	**4.21**

orders (about 250 coefficients). As Table 5.6 show, needlets SN(j_{max}=3) outperforms other representation bases when regressing with both L2 loss and spherical transport loss. Compared with spherical Gaussian and spherical harmonics functions, needlets enable representation in both spatial and frequency domains, thus achieving accurate regression in both domains. Compared with other wavelets such as Haar, needlets is a new generation of spherical wavelets and is more suitable for the representation of spherical image with sharp local peak valleys which are commonly presented in HDR illumination map. Regressing spherical harmonic coefficients with spherical transport loss doesn't improve the performance as there is no spatial information in spherical harmonic coefficients. For spherical Gaussian and Haar, the performance with the spherical transport loss included is clearly improved as the corresponding coefficients are spatially localized, which demonstrates the effectiveness of the proposed spherical transport loss. Further, the prediction performance drops when j_{max}=1,2 are used, and increasing the order to j_{max}=4 doesn't improve the performance. We conjecture that the larger number of parameters with j_{max}=4 affects the regression accuracy negatively. Besides, the performance of SN(j_{max}=3)+ST outperforms SN(j_{max}=3)+HT, demonstrating the superiority of the derived soft thresholding function.

5.5 Conclusion

We designed EMLight, a lighting estimation framework that formulates the illumination prediction as a light distribution regression problem over a spherical surface. A spherical mover's loss is proposed to achieve the effective regression of spherical light distribution. To generate illumination maps with realistic frequency information, we introduce a novel neural projector with spherical convolution that generates panoramic illumination maps through adversarial learning. Quantitative and qualitative experiments show that EMLight is capable of predicting illumination accurately from a single indoor image and rendering the embedded 3D objects realistically. We will continue to investigate illumination estimation from the perspective of spherical distributions in our future works.

We also designed NeedleLight, a lighting estimation model that introduces needlet basis for illumination representation and prediction. In the proposed NeedleLight, we deduce an optimal thresholding function from Bayesian framework for a sparse representation in terms of needlet basis. To tackle the regression of needlet coefficients with spatial localization, a novel spherical transport loss with auxiliary points is designed which performs regression by minimizing the discrepancy between two spherical distributions. Both quantitative and qualitative experiments show that our proposed NeedleLight is capable of predicting illumination accurately from a single indoor image and rendering the embedded 3D objects realistically. We will continue to investigate needlet basis for more efficient illumination representation and explore optimal transport for better network training.

Chapter 6

Conclusion & Future Works

6.1 Conclusion

During my Ph.D. study, I have conducted extensive research into image synthesis in visual machine learning to minimize the gap between synthetic data and real data. Specifically, I have developed a set of image synthesis algorithms for different computer vision tasks, including image synthesis for training data, image synthesis with geometry and appearance realism, and illumination-aware image synthesis as described in the previous three chapters.

In detail, we present composition-based image synthesis in Chapter 3, where we first investigated traditional method for training data synthesis to boost the performance of text detection and recognition. Thanks to the complicated framework of traditional method, we further explored learning-based method with generative adversarial networks which achieves image composition in an end-to-end manner. As current works ignore the potential shadow associated with the foreground object, we proposed to compose images with an auxiliary illumination to guide the generation of realistic shadow effect.

However, composition-based methods require lots of manual setting to achieve best synthetic performance. We thus present translation-based methods in Chapter 4 which achieves automatic image synthesis without manual setting. Our image translation methods can translate images in geometry domain, appearance domain, or both domain for achieving optimal translation realism. Specially, we design a

translation model in geometry domain to rectify curved or perspective scene texts for better text recognition performance. As image translation is naturally a diverse mapping, we proposed a diverse image translation framework in spatial domain to serve as a data augmentation method. Experiments on deep network training and user study demonstrate the superior quality of the synthesized images.

Above methods achieve image synthesis in 2D space, where the image quality is constrained by the limited field of view of 2D space. In Chapter 5, we thus present 3D-level image synthesis with realistic illuminations. As the core of 3D-level image synthesis lies in the prediction of scene illumination, We first explore to generate illumination map through adversarial learning in spatial domain. However, this method ignores the frequency information of illumination which plays an key role in image rendering. So we further introduce a new Needlet basis for illumination representation which enables lighting estimation in both spatial and frequency domain.

6.2 Future Works

We have developed a series of innovative image synthesis techniques that can generate new and realistic images in different approaches. Extensive experiments and analysis have demonstrated that our synthesized images have good potential to be useful in different tasks such as AI for designs, deep network training, etc. On the other hand, image synthesis is still facing a number of challenges before it could be adopted in a wide spectrum of tasks and applications.

In particular, most existing synthesis networks require a large amount of training data [25] which are often laborious to collect. Few-shot synthesis or small-data synthesis [256] with much less reference images can address this constrain effectively. In addition, most existing synthesis networks tend to generate a certain amount of undesired images due to the weak supervision during the image synthesis processes. Novel metrics could be designed to identify high-quality synthesis samples automatically without requiring tedious manual selection of good synthesis samples.

The second is to enhance the controllability and flexibility of the image generation processes and improve the diversity of the contents of the generated images. Most

existing synthesis networks such as CycleGAN [113] leverage adversarial learning with very limited controls in the image generation and translation process which often leads to various artefacts and missing details in the synthesized images. Optimal style control could mitigate this issue in generating high-fidelity and realistic images. In addition, most existing image synthesis networks have very limited diversity in their generated visual features and representation, e.g. the direct image generation from noises produces little new information. This problem could be alleviated by disentangling images into multiple facets and generating them from multiple different data sources that have different data distributions.

The third is to generate visual data in different new dimensions. One new dimension is temporal which generates videos that consist of multiple continuous image frames. Different from generation of static images [27], video generation [257] needs to tackle several new challenges such as smooth transition across the generated neighboring video frames. Another new dimension is spatial which generates 3D visual data that have depth information. It has a wide range of potential applications such as computer games and autonomous driving by generating realistic virtual environments, virtual objects, 3D point clouds, etc.

We will investigate these challenges in our future work.

List of Author's Publications[1]

Publications

- **Fangneng Zhan**, Shijian Lu, Chuhui Xue, "Verisimilar Image Synthesis For Accurate Detection and Recognition of Texts in Scenes", *in European Conference on Computer Vision (ECCV), 2018.*

- **Fangneng Zhan**, Hongyuan Zhu, Shijian Lu, "Spatial Fusion GAN for Image Synthesis", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.*

- **Fangneng Zhan**, Shijian Lu, "ESIR: End-to-end Scene Text Recognition via Iterative Image Rectification", *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.*

- **Fangneng Zhan**, Chuhui Xue, Shijian Lu, "GA-DAN: Geometry-Aware Domain Adaptation Network for Scene Text Detection and Recognition", *in IEEE International Conference on Computer Vision (ICCV), 2019.*

- **Fangneng Zhan**, Shijian Lu, Changgong Zhang, Feiying Ma, Xuansong Xie, " Adversarial Image Composition with Auxiliary Illumination", *in Asia Conference on Computer Vision (ACCV), 2020 (Oral Presentation).*

- **Fangneng Zhan***, Changgong Zhang*, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, Xuansong Xie, "EMLight: Lighting Estimation via Spherical Distribution Approximation", *in AAAI Conference on Artificial Intelligence (AAAI), 2021.*

[1]The superscript * indicates joint first authors

- **Fangneng Zhan**, Changgong Zhang, Shijian Lu, Feiying Ma, Xuansong Xie, "Sparse Needlets for Lighting Estimation with Spherical Transport Loss", *Under Submission*.

Bibliography

[1] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. xvi, 16, 18, 19, 21, 43, 44, 46, 66, 82, 89

[2] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. xx, 11, 94, 95, 103, 104, 105, 106, 108, 109, 114, 116, 117, 118, 119

[3] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7175–7183, 2019. xx, 11, 12, 94, 95, 99, 102, 103, 104, 105, 106, 108, 109, 116, 117, 118

[4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL `http://www.blender.org`. xxi, 104, 116, 119

[5] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016. xxiii, xxiv, 13, 20, 24, 30, 31, 32, 33, 43, 44, 73, 87

[6] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. xxiii, 2, 13, 33, 73

[7] Turner Whitted. An improved illumination model for shaded display. *Communications*, 1980. 1, 9

[8] Henri Gouraud. Continuous shading of curved surfaces. *IEEE transactions on computers*, 100(6):623–629, 1971. 1, 9

[9] James F Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pages 192–198, 1977. 1, 9

[10] David J Heeger. Model for the extraction of image flow. *JOSA A*, 4(8): 1455–1471, 1987. 2

[11] John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.

[12] Brendan McCane, Kevin Novins, D Crannitch, and Ben Galvin. On benchmarking optical flow. *Computer Vision and Image Understanding*, 84(1): 126–143, 2001. 2

[13] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011. 2, 11, 20, 95

[14] Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093, 2012. 2, 13

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2, 13

[16] Ilker Yildirim, Tejas D Kulkarni, Winrich A Freiwald, and Joshua B Tenenbaum. Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual conference of the cognitive science society*, 2015. 2, 13, 20

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 15, 18, 21, 63, 64, 66

[18] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 2, 16, 21, 66

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2, 66

[20] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29:2172–2180, 2016. 2, 16, 21

[21] SC Martin Arjovsky and Leon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia*, 2017. 2, 16, 21, 41, 54

[22] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2, 16, 66, 91

[23] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[24] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3, 16, 36, 66, 103, 124

[26] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3, 16, 102, 125

[28] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10531–10540, 2019. 3

[29] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.

[30] Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, and Jingjing Liu. Bachgan: High-resolution image synthesis from salient object layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[31] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017. 3

[32] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 3

[33] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 4, 19, 34

[34] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Adversarial image composition with auxiliary illumination. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 5, 19, 47

[35] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019. 5, 64, 66

[36] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9105–9115, 2019. 6, 64, 79

[37] Fangneng Zhan, Changgong Zhang, Yu Yingchen, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 7, 94, 95, 96

[38] Fangneng Zhan, Changgong Zhang, Shijian Lu, Feiying Ma, and Xuansong Xie. Sparse needlets for lighting estimation with spherical transport loss. In *Under Submission*, 2021. 7, 94, 108

[39] Apostolia Tsirikoglou, Joel Kronander, Magnus Wrenninge, and Jonas Unger. Procedural modeling and physically based rendering for synthetic data generation in automotive applications. *arXiv preprint arXiv:1710.06270*, 2017. 10

[40] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 10

[41] Subrahmanyan Chandrasekhar. *Radiative transfer*. Courier Corporation, 2013. 10

[42] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 10

[43] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH 2008 classes*, pages 1–10, 2008. 11, 95

[44] Thomas Porter and Tom Duff. Compositing digital images. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 253–259, 1984. 11

[45] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318, 2003. 11, 16, 20, 21

[46] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 11, 19, 20, 57

[47] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep painterly harmonization. In *Computer Graphics Forum*. Wiley Online Library, 2018. 11, 19, 20

[48] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, 2000. 11

[49] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 654–661. IEEE, 2005. 11

[50] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, 98(2):123–145, 2012. 11, 95

[51] Daniel Zoran, Dilip Krishnan, Jose Bento, and Bill Freeman. Shape and illumination from shading using the generic viewpoint assumption. In *Advances in Neural Information Processing Systems*, pages 226–234, 2014.

[52] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 11

[53] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2017.

[54] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4080–4089, 2019.

[55] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2020. 11, 95

[56] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8598–8607, 2019. 11

[57] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 38 (1):129–141, 2015. 11, 95

[58] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3114–3122, 2017. 11, 95

[59] Dachuan Cheng, Jian Shi, Yanyun Chen, Xiaoming Deng, and Xiaopeng Zhang. Learning scene illumination by pairwise photos from rear and front mobile cameras. In *Computer Graphics Forum*, pages 213–221. Wiley Online Library, 2018. 11, 50, 94, 95, 108

[60] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 11, 94, 95

[61] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 11, 12, 50, 94, 95, 103, 104, 105, 106, 107, 108, 116, 117, 118

[62] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019. 11, 12, 94, 95, 108

[63] Mengtian Li, Jie Guo, Xiufen Cui, Rui Pan, Yanwen Guo, Chenchen Wang, Piaopiao Yu, and Fei Pan. Deep spherical gaussian illumination estimation for indoor scene. In *Proceedings of the ACM Multimedia Asia*, pages 1–6, 2019. 12, 95, 104, 105, 106, 116, 117, 118, 119

[64] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for

unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019. 12, 95, 103, 104, 116

[65] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 12, 94, 95, 108

[66] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Teaching 3d geometry to deformable part models. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3362–3369. IEEE, 2012. 12

[67] Scott Satkin, Jason Lin, and Martial Hebert. Data-driven scene understanding from 3d models. In *BMVC*, 2012. 12

[68] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, pages 640–653. Springer, 2012. 12

[69] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. In *null*, page 641. IEEE, 2003. 12

[70] Vikash K Mansinghka, Tejas D Kulkarni, Yura N Perov, and Josh Tenenbaum. Approximate bayesian image interpretation using generative probabilistic graphics programs. In *Advances in Neural Information Processing Systems*, pages 1520–1528, 2013. 12

[71] David Vazquez, Antonio M Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. Virtual and real world adaptation for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):797–809, 2013. 12

[72] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 12

[73] Geoffrey R Taylor, Andrew J Chosak, and Paul C Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 13

[74] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014. 13

[75] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1278–1286, 2015. 13

[76] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 13

[77] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642, 2015. 13

[78] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2821–2840, 2012. 13

[79] Joerg Liebelt, Cordelia Schmid, and Klaus Schertler. independent object class detection using 3d feature maps. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 13

[80] Francisco Massa, Bryan C Russell, and Mathieu Aubry. Deep exemplar 2d-3d detection by adapting from real to rendered views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6024–6033, 2016. 13

[81] Stephan R Richter and Stefan Roth. Discriminative shape from shading in uncalibrated illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1128–1136, 2015. 13

[82] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 13

[83] Biliana Kaneva, Antonio Torralba, and William T Freeman. Evaluation of image features using a photorealistic virtual world. In *2011 International Conference on Computer Vision*, pages 2282–2289. IEEE, 2011. 13

[84] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. 13

[85] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2722–2730, 2015. 13

[86] Shangbang Long and Cong Yao. Unrealtext: Synthesizing realistic scene text images from the unreal world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5488–5497, 2020. 13

[87] Xiao Yang, Dafang He, Daniel Kifer, and C Lee Giles. A learning-based text synthesis engine for scene text detection. In *30th British Machine Vision Conference, BMVC 2019*, 2020. 13

[88] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 167:50–62, 2018. 13

[89] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 189–205, 2018. 13

[90] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 608–617, 2019. 14

[91] Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu. Attribute analysis with synthetic dataset for person re-identification. *arXiv preprint arXiv:2006.07139*, 2020. 14

[92] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007. 14

[93] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455, 2009. 14

[94] Charles J Geyer. Markov chain monte carlo maximum likelihood. In *Computing science and statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation, Fairfax Station, 1991*, pages 156–163, 1991. 14

[95] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 14

[96] Asja Fischer and Christian Igel. An introduction to restricted boltzmann machines. In *Iberoamerican congress on pattern recognition*, pages 14–36. Springer, 2012. 14

[97] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6:3–10, 1993. 15

[98] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 15

[99] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153–160, 2006. 15

[100] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 15

[101] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27:3581–3589, 2014. 15

[102] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in neural information processing systems*, pages 899–907, 2013. 15

[103] Yoshua Bengio, Eric Laufer, Guillaume Alain, and Jason Yosinski. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pages 226–234, 2014. 15

[104] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 15

[105] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 15

[106] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 15

[107] Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. *arXiv preprint arXiv:1511.06406*, 2015. 15

[108] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017. 15

[109] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 15

[110] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 16, 66

[111] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 16

[112] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018. 16, 21

[113] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 16, 36, 66, 82, 89, 91, 125

[114] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 16

[115] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[116] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018.

[117] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 16

[118] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *arXiv preprint arXiv:1703.01560*, 2017. 16, 21

[119] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. *Advances in neural information processing systems*, 31:10393–10403, 2018. 16, 21

[120] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2487–2495, 2019. 16, 21

[121] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3943–3951, 2015. 16, 21

[122] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *arXiv preprint arXiv:1807.07560*, 2018. 16, 18, 21

[123] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2019. 16, 21, 56, 57

[124] R. Timofte, R. Rothe, and L. V. Gool. Seven ways to improve example-based single image super resolution. *CVPR*, 2016. 17

[125] Donald P Greenberg, Kenneth E Torrance, Peter Shirley, James Arvo, Eric Lafortune, James A Ferwerda, Bruce Walter, Ben Trumbore, Sumanta Pattanaik, and Sing-Choong Foo. A framework for realistic image synthesis. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 477–494, 1997. 18

[126] G. Papandreou, L. C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. *International Conference on Computer Vision (ICCV)*, pages 1742–1750, 2015. 18

[127] S. Tian and S. Lu. Wetext: Scene text detection underweak supervision. *IEEE International Conference on Computer Vision*, pages 1492–1550, 2017. 18

[128] Leon Sixt, Benjamin Wild, and Tim Landgraf. Rendergan: Generating realistic labeled data. *arXiv:1611.01331*, 2017. 18

[129] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018. 19, 20, 22, 43, 44

[130] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Towards realistic 3d embedding via view alignment. *arXiv preprint arXiv:2007.07066*, 2020. 19

[131] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. *Proc. ICCV*, 2015. 20

[132] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Scene text synthesis for efficient and effective deep network training. *arXiv preprint arXiv:1901.09193*, 2019. 20

[133] Chuhui Xue, Shijian Lu, and Fangneng Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–372, 2018. 20

[134] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014. 20, 33, 43, 44, 45, 64, 75

[135] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017. 20

[136] Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Guided image inpainting: Replacing an image region by pulling content from another image. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1514–1523. IEEE, 2019. 20

[137] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. *arXiv preprint arXiv:2104.12335*, 2021. 20

[138] Matthew Uyttendaele, Ashley Eden, and Richard Skeliski. Eliminating ghosting and exposure artifacts in image mosaics. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001. 20

[139] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010. 20

[140] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012. 20

[141] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. In *European Conference on Computer Vision*, pages 31–44. Springer, 2010. 20

[142] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, and Ming-Hsuan Yang. Sky is not the limit: semantic-aware sky replacement. *ACM Trans. Graph.*, 35(4):149–1, 2016. 20

[143] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 21

[144] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020. 21

[145] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 21

[146] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8139–8148, 2020. 21, 95

[147] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *European conference on computer vision*, pages 1–14. Springer, 2010. 22, 51

[148] Kaiming He and Jian Sun. Fast guided filter. *arXiv preprint arXiv:1505.00996*, 2015. 22

[149] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013. 22

[150] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. Rolling guidance filter. In *European conference on computer vision*, pages 815–830. Springer, 2014. 22

[151] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2018. 22

[152] Wei Liu, Yun Gu, Chunhua Shen, Xiaogang Chen, Qiang Wu, and Jie Yang. Data driven robust image guided depth map restoration. *arXiv preprint arXiv:1512.08103*, 2015. 22

[153] Wei Liu, Xiaogang Chen, Chunhua Shen, Jingyi Yu, Qiang Wu, and Jie Yang. Robust guided image filtering. *arXiv preprint arXiv:1703.09379*, 2017. 22

[154] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust guided image filtering using nonconvex potentials. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):192–207, 2017. 22

[155] David R Hunter and Kenneth Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000. 22

[156] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 24, 25

[157] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013. 25

[158] Shijian Lu, Cheston Tan, and Joo-Hwee Lim. Robust and efficient saliency modeling from image co-occurrence histograms. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):195–201, 2013. 25

[159] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013. 27, 29, 30, 42, 43, 44, 73, 86, 89

[160] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. *arXiv:1704.03155*, 2017. 28, 30, 87

[161] K. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. Pvanet: Deep but lightweight neural networks for real-time object detection. *arXiv:1608.08021*, 2016. 28

[162] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. 28, 33, 64, 65, 67, 72, 75, 76, 77, 78

[163] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 29, 30, 42, 43, 44, 73, 86

[164] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. *Computer Vision and Pattern Recognition (CVPR)*, pages 1083–1090, 2012. 29, 30

[165] Jerod J Weinman, Erik Learned-Miller, and Allen R Hanson. Scene text recognition using similarity and a lexicon with sparse belief propagation.

IEEE Transactions on pattern analysis and machine intelligence, 31(10): 1733–1746, 2009. 29, 42, 44, 73, 86, 89

[166] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011. 29, 42, 43, 44, 74, 86, 89

[167] Christian Wolf and Jean-Michel Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(4):280–296, 2006. 29

[168] L. Kang, Y. Li, and D. Doermann. Orientation robust textline detection in natural images. *In Proc. of CVPR*, 2014. 30

[169] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao. Multi-orientation scene text detection with adaptive clustering. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1930–1937, 2015. 30

[170] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116(1):1–20, 2016. 30, 75, 76

[171] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks. *In Proc. of CVPR*, 2015. 30

[172] Z. Tian, W. Huang, P. He T. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. *European Conference on Computer Vision*, pages 56–72, 2016. 30

[173] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao. Scene text detection via holistic, multi-channel prediction. *arXiv:1606.09002*, 2016. 30

[174] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012. 33, 64

[175] Jose A Rodriguez-Serrano, Albert Gordo, and Florent Perronnin. Label embedding: A frugal baseline for text recognition. *International Journal of Computer Vision*, 113(3):193–207, 2015. 33, 64

[176] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4049, 2014. 33

[177] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014. 33

[178] Albert Gordo. Supervised mid-level features for word image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2956–2964, 2015. 33

[179] Alessandro Bissacco, Mark Cummins, Yuval Netzer, and Hartmut Neven. Photoocr: Reading text in uncontrolled conditions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 785–792, 2013. 33

[180] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016. 33

[181] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 36, 66, 91

[182] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30: 700–708, 2017. 36, 66, 91

[183] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989. 36, 46, 70, 83

[184] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–576, 2013. 43, 44, 65, 67, 74, 87, 89

[185] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 43, 44, 74, 87, 89

[186] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 43, 46

[187] Bolan Su and Shijian Lu. Accurate scene text recognition based on recurrent neural network. In *Asian Conference on Computer Vision*, pages 35–48. Springer, 2014. 44

[188] Bolan Su and Shijian Lu. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recognition*, 63:397–405, 2017. 64

[189] Shangxuan Tian, Ujjwal Bhattacharya, Shijian Lu, Bolan Su, Qingqing Wang, Xiaohua Wei, Yue Lu, and Chew Lim Tan. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. *Pattern Recognition*, 51:125–134, 2016.

[190] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 65, 67, 72, 73, 74, 75, 76, 77, 78

[191] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5571–5579, 2018. 44, 73, 75

[192] Robin Green. Spherical harmonic lighting: The gritty details. In *Archives of the Game Developers Conference*, volume 56, page 4, 2003. 49

[193] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 49, 72

[194] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 55

[195] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. 55

[196] Liming Wang, Jianbo Shi, Gang Song, and I-Fan Shen. Object detection combining recognition and segmentation. In *Asian conference on computer vision*, pages 189–199. Springer, 2007. 55

[197] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint arXiv:1511.04136*, 2015. 55, 56

[198] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 55, 56

[199] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su,

et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 55, 56

[200] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30:6626–6637, 2017. 56

[201] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 56

[202] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 56

[203] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1053–1061, 2018. 57

[204] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 58

[205] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 63, 67

[206] Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Edit probability for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1508–1516, 2018. 64

[207] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Synthetically supervised feature learning for scene text recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 64

[208] Shijian Lu, Ben Mei Chen, and Chi Chung Ko. Perspective rectification of document images using fuzzy set and morphological operations. *Image and Vision Computing*, pages 541–553, 2005. 65

[209] Paul Clark and Majid Mirmehdi. Recognising text in real scenes. In *IJDAR*, pages 243–257, 2002.

[210] Shijian Lu, Ben M Chen, and Chi Chung Ko. A partition approach for the restoration of camera images of planar and curled document. *Image and Vision Computing*, 24(8):837–848, 2006. 65

[211] Christian Bartz, Haojin Yang, and Christoph Meinel. See: towards semi-supervised end-to-end scene text recognition. *arXiv preprint arXiv:1712.05404*, 2017. 65

[212] Wei Liu, Chaofeng Chen, and Kwan-Yee K Wong. Char-net: A character-aware neural network for distorted scene text recognition. In *AAAI*, 2018. 65

[213] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 65

[214] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 65

[215] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 65

[216] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015. 66

[217] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 66

[218] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *arXiv preprint arXiv:1608.06019*, 2016. 66

[219] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 66

[220] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 66

[221] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 66

[222] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 66

[223] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 66

[224] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 66, 91

[225] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 66

[226] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 72

[227] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084, 2017. 74, 75, 76

[228] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. Learning to read irregular text with attention mechanisms. In *IJCAI*, 2017. 75

[229] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012. 86

[230] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5909–5918, 2018. 87

[231] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *ECCV*, pages 20–36, 2018. 87, 88

[232] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. 88

[233] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 90

[234] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in neural information processing systems*, 29:469–477, 2016. 91

[235] Zhang Chen, Anpei Chen, Guli Zhang, Chengyuan Wang, Yu Ji, Kiriakos N Kutulakos, and Jingyi Yu. A neural rendering framework for free-viewpoint relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5599–5610, 2020. 94

[236] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Changgong Zhang, Shijian Lu, Ling Shao, Feiying Ma, and Xuansong Xie. Gmlight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*, 2021. 95

[237] Edward Zhang, Michael F Cohen, and Brian Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6): 1–14, 2016. 95

[238] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4): 79–1, 2019. 95

[239] Peter Schröder and Wim Sweldens. Spherical wavelets: Efficiently representing functions on the sphere. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 161–172, 1995. 96

[240] Francis J Narcowich, Pencho Petrushev, and Joseph D Ward. Localized tight frames on spheres. *SIAM Journal on Mathematical Analysis*, 38(2):574–594, 2006. 96, 110

[241] Paolo Baldi, Gérard Kerkyacharian, Domenico Marinucci, Dominique Picard, et al. Asymptotics for spherical needlets. *The Annals of Statistics*, 37(3): 1150–1171, 2009. 96, 110

[242] Davide Pietrobon, Amedeo Balbi, and Domenico Marinucci. Integrated sachs-wolfe effect from the cross correlation of wmap3 year and the nrao vla sky survey data: New results and constraints on dark energy. *Physical Review*, 2006. 96

[243] D Marinucci, D Pietrobon, A Balbi, P Baldi, P Cabella, G Kerkyacharian, Paolo Natoli, D Picard, and N Vittorio. Spherical needlets for cosmic microwave background data analysis. *Monthly Notices of the Royal Astronomical Society*, 383(2):539–545, 2008. 96

[244] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 98

[245] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–533, 2018. 98, 102

[246] Helmut Vogel. A better way to construct the sunflower head. *Mathematical biosciences*, 44(3-4):179–189, 1979. 99, 101, 102

[247] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 101, 114

[248] Jérome Bobin, Yassir Moudden, J-L Starck, J Fadili, and Nabila Aghanim. Sz and cmb reconstruction using generalized morphological component analysis. *Statistical Methodology*, 5(4):307–317, 2008. 112

[249] Ron O Dror, Alan S Willsky, and Edward H Adelson. Statistical characterization of real-world illumination. *Journal of Vision*, 4(9):11–11, 2004. 112

[250] Davide Maino, A Farusi, Carlo Baccigalupi, Francesca Perrotta, AJ Banday, L Bedini, Carlo Burigana, Gianfranco De Zotti, KM Górski, and E Salerno. All-sky astrophysical component separation with fast independent component analysis (fastica). *Monthly Notices of the Royal Astronomical Society*, 334(1):53–68, 2002. 112

[251] Flavien Vansyngel, Benjamin D Wandelt, Jean-François Cardoso, and Karim Benabed. Semi-blind bayesian inference of cmb map and power spectrum. *Astronomy & Astrophysics*, 588:A113, 2016. 112

[252] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018. 113

[253] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016. 113, 114

[254] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 880–889. PMLR, 2018. 115

[255] Colm Mulcahy. Image compression using the haar wavelet transform. *Spelman Science and Mathematics Journal*, 1(1):22–31, 1997. 120

[256] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 124

[257] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 125