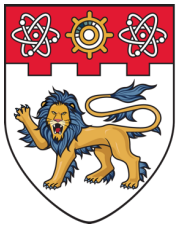
Visual Saliency Computation and Quality Evaluation via Deep Learning



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Sheng YANG

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

March 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.

27-Mar-2020
.....
Date

Yang Sheng
.....
Sheng YANG

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.

27-Mar-2020

.....

Date

.....

Prof. Weisi Lin

Authorship Attribution Statement

This thesis contains material from 3 papers published or submitted in the following peer-reviewed journal and conferences in which I am the first author.

Chapter 3 is published as S. Yang, G. Lin, Q. Jiang, W. Lin, A Dilated Inception Network for Visual Saliency Prediction. IEEE Transactions on Multimedia (TMM), 2019. DOI:10.1109/TMM.2019.2947352.

The contributions of the co-authors are as follows:

- Prof. Weisi Lin and Prof. Guosheng Lin supervised me on the overall research direction and edited the manuscript drafts.
- I designed the study, performed the experiments, and prepared the manuscript drafts.
- Prof. Qiuping Jiang assisted in the experiments and revised the manuscript drafts.

Chapter 4 is submitted as S. Yang, W. Lin, G. Lin, Z. Liu, Q. Jiang, Progressive Self-Guided Loss for Salient Object Detection. In Submission.

The contributions of the co-authors are as follows:

- Prof. Weisi Lin supervised me on the overall research direction and edited the manuscript drafts.
- I proposed the idea, performed the experiments, and prepared the manuscript drafts.
- Prof. Guosheng Lin helped me in designing the algorithm and reorganizing the content of the manuscript drafts.
- I and Dr. Zichuan Liu co-designed the experiments.
- Prof. Qiuping Jiang revised the manuscript drafts.

Chapter 5 is published as S. Yang, Q. Jiang, W. Lin, Y. Wang, SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment. In Proceedings of the 27th ACM International Conference on Multimedia (MM '19), 2019. DOI: 10.1145/3343031.3350990.

The contributions of the co-authors are as follows:

- Prof. Weisi Lin and Prof. Yongtao Wang supervised me on the overall research direction and edited the manuscript drafts.
- I proposed the idea, performed the experiments, and prepared the manuscript drafts.

- I and Prof. Qiuping Jiang co-designed and revised the content of the manuscript drafts.

27-Mar-2020 Yang Sheng

....

Date Sheng YANG

Acknowledgements

First of all, I would like to express my greatest gratitude to my supervisor, Prof. Weisi Lin, for his insightful guidance, continuous support, and enthusiastic encouragement during my Ph.D. study. I still remember those tough days that I was struggling with my research. Without his patience and guidance, I would not be able to complete my first research work. His inspiring and down-to-earth working attitude has greatly influenced my academic career.

Special thanks to my collaborators, Prof. Guosheng Lin, Prof. Yongtao Wang, Prof. Qiuping Jiang, and Dr. Zichuan Liu for their invaluable suggestions and continuous encouragement on my research works.

I really enjoyed the days at Nanyang Technological University. I wish to thank my teammates for their sincere friendship and interesting discussions: Yuan Yuan, Yabin Zhang, Qiaohong Li, Wentao Cheng, Shasha Mao, Ke Gu, Guanghui Yue, Yupeng Cheng, Zhuo Chen, Baoquan Zhao, Jong-Uk Hou, Chenlei Lv, Jian Jin, Guoqing Zhang, and Jingwen Hou. I always appreciate the precious moments we shared together. I also thank my colleagues and technicians in Media & Interactive Computing Lab and ST Engineering-NTU Corporate Lab for their continuous support and help. I would also like to express my thanks to my lovely friends, Yitao Han, Ao Zhou, Panrong Tong, Xu Yang, Chi Zhang, Xiaofeng Yang, and those who are not listed here. I am very happy to meet you in Singapore.

Last but not least, I would like to thank my dear family for their unconditional love and warm support throughout my life. I wish you would be proud of me.

Abstract

Visual attention is an important mechanism in our human vision system, which filters out redundant and unimportant visual information for selectively processing the most salient or informative regions from the visual field. Visual saliency computation is about understanding and simulating the behavior of this selective attention mechanism in a visual scene. Computational models for visual saliency can provide clues to where people will look in images, what objects are salient in a scene, and how people will evaluate the perceptual quality of an image. Such models can be applied to advance a wide range of visual-oriented applications in image processing and computer vision areas. At present, recent advances in visual saliency computation are mainly led by the progress in deep learning techniques and many deep learning-based visual saliency approaches have emerged.

In this thesis, we study the problems of deep learning-based visual saliency computation, including saliency prediction and salient object detection (SOD). Besides, saliency-guided image quality evaluation is also investigated to extend our work. For saliency prediction, existing deep saliency models suffer from either huge computation cost or limited performance gain. We propose an effective yet efficient saliency model, named Dilated Inception Network (DINet), to characterize the diverse and effective saliency-influential factors at different receptive field sizes with much smaller computation cost. Experimental results on the challenging saliency prediction datasets demonstrate the outstanding performance of our model in terms of both speed and accuracy.

For SOD, the saliency maps produced by previous works still suffer from incomplete predictions due to the internal complexity of salient objects. To alleviate this problem, we propose a simple yet effective progressive self-guided loss function (PSG loss) to create progressive and auxiliary training supervisions for step-wisely guiding the training process. In our PSG loss, a simulated morphological closing operation is applied to the network predictions to generate the needed progressive

supervisions epoch-wisely for characterizing the spatial dependencies of salient object pixels. As a result, SOD models can be guided by these generated supervisions to highlight more complete salient objects step-by-step for alleviating the problem of incomplete predictions. Experimental results on six widely used SOD benchmark datasets show that our loss function not only advances the performance of existing SOD models without architecture modification but also helps our proposed framework to achieve state-of-the-art performance.

In the last work, we propose a novel saliency-guided deep neural network (SGDNet) to incorporate learnable saliency information into image quality evaluation. This model is the first attempt to jointly optimize the saliency prediction and quality evaluation sub-tasks in an end-to-end multi-task learning framework. The learned saliency information from the saliency prediction sub-task is transparent to the quality evaluation sub-task by providing a kind of spatial attention priors for the perceptually-consistent feature fusion. The effectiveness of the learned saliency information and the proposed multi-task framework are validated in the experiments.

Contents

List of Figures

List of Tables

List of Abbreviations

ASPP	Atrous Spatial Pyramid Pooling
AUC	Area Under the ROC Curve
BAM	Branch-wise Attention Mechanism
BCE	Binary Cross Entropy
CA	Channel-wise Attention
CC/PLCC	Pearson Linear Correlation Coefficient
CCE	Categorical Cross Entropy
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DCNN	Deep Convolutional Neural Network
DIM	Dilated Inception Module
DINet	Dilated Inception Network
DRN	Dilated Residual Network
FC	Fully Connected
FCN	Fully Convolutional Network
FOV	Field-of-View
FPN	Feature Pyramid Network
GAN	Generative Adversarial Networks
GAP	Global Average Pooling
GT	Ground Truth
HVS	Human Visual System
IPN	Image Pyramid Network
IQA	Image Quality Assessment
KLD	Kullback–Leibler Divergence
LSTM	Long Short Term Memory
MAE	Mean Average Difference
maxF	Maximum F-measure

MLP	Multi-layer Perceptron
MOS	Mean Opinion Score
MS-FAM	Multi-scale Feature Aggregation Module
NR-IQA	No-reference Image Quality Assessment
NLB	Non-local Block
NSS	Normalized Scanpath Saliency
NSS	Natural Scene Statistics
PD	Probability Distribution
PSG Loss	Progressive Self-Guided Loss
PR	Precision-Recall
ROIs	Regions of Interest
sAUC	Shuffled Area Under the ROC Curve
SE	Squeeze-and-Excitation
SGDNet	Saliency-guided Deep Neural Network
SOD	Salient Object Detection
SRCC	Spearman Rank Order Correlation Coefficient
SVM	Support Vector Machine
TV Distance	Total Variation Distance

Chapter 1

Introduction

1.1 Background and Motivations

Visual attention mechanism refers to the ability of the human visual system (HVS) to automatically select the most salient or informative regions from natural scenes by filtering out redundant and unimportant visual information for further processing. Around 10^8-10^9 bits per second of visual data enters into our eyes, as reported in [13]. Without the help of this visual attention mechanism, the HVS is not able to handle and process this large volume of data in real-time. Therefore, it is important to understand and simulate the behavior of visual attention to advance a wide range of visual-oriented multimedia applications such as image retargeting [14], image and video compression [15, 16], image quality assessment [17–19], video summarization [20], virtual reality content design [21], and more. Motivated by these perception-aware applications, some visual saliency models, which focus on predicting human eye fixations [13], are firstly emerged and developed to facilitate them. With the rising need for many object-level computer vision applications, such as object detection and recognition [22–25], image editing and manipulating [26, 27], visual tracking [28], and semantic segmentation [29], a new type of visual saliency task is created by incorporating the high-level concept of salient objects into the process of visual saliency computation [30]. Saliency models in this new type aim to segment the entire salient foreground objects from the background [31].

Although both types of saliency models are expected to be applied interchangeably, their generated saliency maps are quite different [31]. In particular, fixation

FIGURE 1.1: Sample images and their ground truths for two different visual saliency computation tasks. (a) Sample images. (b) The ground truths for eye fixation prediction. (c) The ground truths for salient object detection.

prediction models usually generate sparse blob-like salient regions for reflecting the fixation distribution, as shown in Fig. 1.1(b), while salient object detection models often highlight smooth connected areas or binary masks for the salient objects, as shown in Fig. 1.1(c). Such a difference demonstrates their different characteristics due to the distinct purposes of visual saliency. Recent studies [32, 33] have explored the relationship between these two types of saliency maps. Saliency maps generated by fixation prediction models are instinctive and more consistent with the visual attention mechanism in natural scene free-viewing. They can be used as a kind of low-level saliency priors to explicitly indicate the regions of interest (ROIs) where may exist the salient objects.

The computational visual saliency models generally follow a two-stage processing pipeline including feature extraction and saliency inference. Before the advent of deep learning, classic saliency methods usually adopt multiple carefully designed hand-crafted features and perform the saliency inference by fusing these features

in a heuristic way [34]. Starting from the success of AlexNet [35] in the image classification competition [36], deep learning has progressively become the dominant technique for computer vision as well as other applications. Not surprisingly, recent advances in both visual saliency prediction and salient object detection are mainly boosted by the progress in deep learning techniques. As a result, the top of various benchmarking leaderboards in these two tasks are occupied by those deep learning-based methods. Compared to conventional approaches, deep learning provides the powerful capability of automatic feature extraction and aggregation from the labeled data and simplifies the processing pipeline by inferring the visual saliency directly from the input image in an end-to-end manner. As such, it eliminates the need for repeatedly designing and testing hand-crafted features, alleviates the heavy dependency on domain knowledge, and hence has been adopted by many researchers and engineers.

In this thesis, we study the problems of deep learning-based visual saliency computation that range from low-level fixation prediction to high-level salient object detection. Besides, saliency-guided image quality evaluation is also investigated as it is a typical application of visual saliency. In the following, we will describe the existing challenges and introduce our motivations for these problems.

For visual saliency prediction, although deep learning-based saliency models have shown their superior performance towards the conventional methods, it can be further improved by fully characterizing the multi-scale saliency-influential factors [1, 37, 38]. Existing deep saliency models [38–40], which contain some powerful or simple multi-scale feature extraction modules, usually suffer from either huge computation cost or limited performance gain, respectively. To get rid of this awkward situation, we propose a new multi-scale feature extraction module, dilated inception module (DIM), to capture more effective and diverse multi-scale contextual features at a much lower computation cost. As such, a better trade-off between the computation cost and performance gain can be achieved.

For salient object detection (SOD), most deep learning-based SOD models use the binary cross-entropy (BCE) as their training loss. But BCE loss is a typical pixel-wise loss function which only accounts for the pixel-wise difference between labels and predictions, ignoring the spatial dependencies of salient object pixels. Models trained with BCE loss usually have the problem of incomplete predictions, as shown in Fig. 4.1, since every pixel is predicted individually. Therefore, a more suitable

training loss is required. Several efforts [41–43] have been made in this direction. However, their proposed losses are still not specifically designed for capturing the spatial dependencies among salient pixels. From another perspective, one may ask: "Can we obtain better performance even trained with the simple BCE loss?" Compared with the efforts of designing more suitable loss, the investigations on the training targets are seldom investigated. In this work, we propose to progressively modify the training supervisions to create a progressive self-guided (PSG) loss for characterizing the spatial dependencies of salient object pixels. As a result, SOD models can be guided step-by-step by these progressive supervisions to highlight more complete salient objects, even trained with the simple BCE loss.

For no-reference image quality assessment (NR-IQA), some recent deep learning-based IQA methods [44, 45] seek to use a multi-task learning strategy for learning a better feature representation by jointly optimizing two sub-networks for two different sub-tasks. However, these methods usually use distortion identification as the auxiliary sub-task, which cannot accurately identify the diverse and complex mixtures of distortions that exist in the real-world images, for the primary quality evaluation sub-task. To solve this limitation, we propose to use visual saliency prediction to replace it as the auxiliary sub-task for providing more universal yet closely related perceptual information in the quality evaluation sub-network. The advantages of our approach are twofold: Firstly, visual saliency always exists when viewing every image, regardless of its distortion type. As such, the inability of existing methods on evaluating the qualities of authentically distorted images is disappeared. Secondly, by incorporating learnable visual saliency information into image quality evaluation via a multi-task learning framework, the visual importance dependence is introduced into the deep IQA models as a kind of adaptive spatial attention priors to further facilitate them. Besides, most of the existing methods [44–47] suffer from the label noise problem caused by assigning the global quality label for all the patches cropped from the same input image. We solve it by implementing an image-based approach that evaluates image quality directly from the whole input image instead of its local patches to avoid this potential problem.

1.2 Objective and Scope

The objective of this thesis is to build new deep learning-based computational models for visual saliency prediction and salient object detection, and explore the application of the proposed saliency models in image quality assessment. In particular, we explore the following problems in this thesis:

- To achieve lower computation cost and larger performance gain in visual saliency prediction.
- To alleviate the problem of incomplete predictions in existing salient object detection methods.
- To address the visual importance dependence and label noise problems in existing image quality assessment methods.

1.3 Summary of Contributions

Firstly, based on the analysis of the limitations of existing deep learning-based visual saliency prediction methods, an effective yet efficient saliency model, named Dilated Inception Network (DINet), is proposed to achieve a better trade-off between the computation cost and performance gain. This model is equipped with our proposed dilated inception module (DIM) to characterize the diverse saliency-influential factors at different receptive field sizes with a much smaller computation cost. The scale diversity of the captured multi-scale contextual features is enriched by introducing paralleled dilated convolutions with various dilation ratios in this effective and lightweight DIM. In particular, a new visualization method specially designed for DIM is proposed to verify its effectiveness. Besides, a set of linear normalization-based probability distribution distance metrics are proposed as loss functions to optimize our DINet. As such, the saliency prediction task is formulated as a probability distribution prediction problem, consequently leading to an extra performance gain. The performance of our DINet is evaluated on various saliency prediction datasets. The peer comparison results indicate that our DINet can achieve state-of-the-art performance in terms of both accuracy and speed.

Secondly, to enhance the representation capacity of the CNN models to explicitly learn the spatial dependencies within the entire salient objects without architecture modification, a simple yet effective progressive self-guided loss function (PSG loss) is proposed to guide the salient object detection (SOD) model to highlight more complete salient objects step-by-step. The proposed PSG loss is the first attempt to create progressive and auxiliary training supervisions for step-wisely guiding the training process. In our PSG loss, an imitated morphological closing operation is applied to the network predictions to generate such progressive training supervisions. As a result, the spatial dependencies of salient object pixels are characterized since these generated supervisions are progressively expanded from their sources. A new multi-scale feature aggregation module, which is composed of our DIM and a branch-wise attention mechanism, is proposed to build our SOD framework for further improvement. Benefiting from this module, our SOD framework takes full advantage of adaptive multi-scale feature aggregation to locate and detect salient objects effectively. The performance of our proposed SOD model is evaluated on six widely used SOD datasets. The peer comparison results indicate that our model can achieve state-of-the-art performance with the help of our proposed PSG loss. Meanwhile, PSG loss can be directly applied to train other existing SOD models without architecture modification for alleviating their incomplete prediction problem and thus advancing their performance.

Thirdly, to overcome the inability of existing multi-task CNN-based no-reference image quality assessment (NR-IQA) methods on evaluating the perceptual qualities of real-world images with authentic distortions, a saliency-guided deep neural network (SGDNet) is proposed to jointly optimize two sub-tasks including visual saliency prediction and image quality evaluation with a shared feature extractor. Related works have reported that saliency information is highly correlated with image quality while this property is fully utilized in our proposed SGNet by training the model with more informative labels including proxy saliency maps and quality scores simultaneously. As such, more discriminant features can be learned and more accurate mapping from feature representations to quality scores can be established. Moreover, our SGDNet model is an image-based approach that generates feature maps from the whole input images instead of their local patches to avoid the potential label noise problem. The performance of our SGNet is evaluated on several publicly available IQA datasets. The peer comparison results indicate that our SGDNet can achieve state-of-the-art performance on both

authentically and synthetically distorted IQA datasets. Meanwhile, the ablation study shows that the quality evaluation performance is indeed boosted by incorporating saliency information and our multi-task learning framework can further improve the performance due to its learned adaptive spatial attention priors for better perceptually-consistent feature fusion.

1.4 Outline of the Thesis

The thesis consists of six chapters and each chapter is summarized as follows. Chapter 1 gives a brief introduction to the background and motivations of this thesis, and summarizes the objective and contributions. Chapter 2 reviews the related works in the literature. Chapter 3 presents our first work which proposes a dilated inception network (DINet) for visual saliency prediction. Chapter 4 focuses on the second work which proposes a progressive self-guided loss (PSG loss) for salient object detection. Chapter 5 describes a saliency-guided deep neural network (SGDNet) for no-reference image quality assessment. The thesis is concluded in Chapter 6 and the potential future work is discussed.

Chapter 2

Literature Review

In this chapter, we review the previous works that are related to the research topics covered in this thesis. In Section 2.1, we review the saliency models for visual saliency prediction. In Section 2.2, we present other saliency models for salient object detection. In the last section, we discuss the no-reference image quality assessment (NR-IQA) methods and explore the relationship between visual saliency and NR-IQA. An overview of the related work and its relationship to our research is depicted in Fig. 2.1.

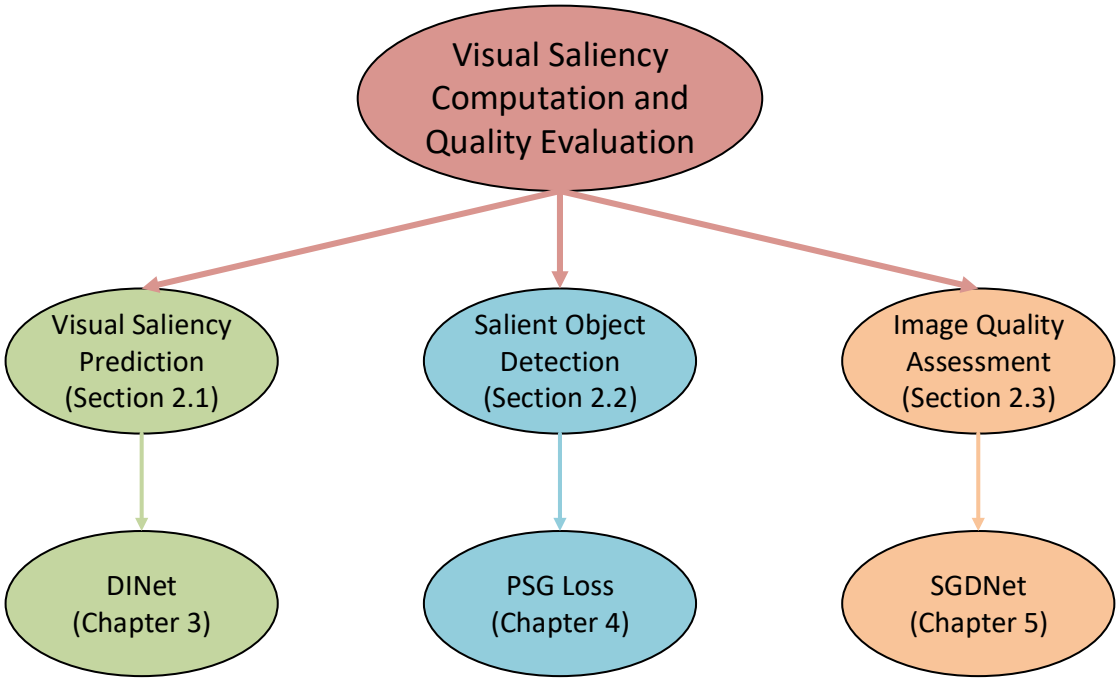


FIGURE 2.1: An overview of the related work and its relationship to our research.

2.1 Visual Saliency Prediction

Visual saliency prediction, also known as eye fixation prediction, aims to compute a saliency map that topographically represents humans' attentional priority when they freely view a given image [48]. In this section, we first review the classic saliency prediction methods. Then, we present recent advances in saliency prediction by utilizing some emerging deep learning techniques. Finally, we summarize the existing deep saliency models with multi-scale feature extraction modules for showing our motivation for this task.

2.1.1 Visual Saliency Prediction Methods before the advent of Deep Learning

The classic saliency prediction methods can be roughly classified into three typical categories: i) biologically plausible saliency methods; ii) spectral analysis-based saliency methods; and iii) machine learning-based saliency methods.

2.1.1.1 Biologically Plausible Saliency Methods

Most of classic saliency prediction models [49–51] are biologically plausible. They mainly extract multiple low-level hand-crafted features, such as intensity, color, and so on, and fuse these features in a heuristic way.

One of the seminal works in visual saliency was proposed by Itti *et al.* [49]. In their framework, the feature maps are obtained by calculating multi-scale center-surround differences for color, intensity, and orientation. The final saliency map is achieved by fusing these feature maps via a linear combination. Later, Harel *et al.* proposed a Graph-Based Visual Saliency (GBVS) model in [50]. They follow Itti's approach to extract feature maps. Then, a fully connected graph is built for covering all grid locations of each feature map. In these graphs, weights between two nodes are assigned proportionally to their feature similarity and spatial distance for measuring their visual closeness. The resulting graphs are treated as Markov chains and their equilibrium distribution is further normalized and combined into the final saliency maps.

Some of the biologically plausible saliency models are based on the assumption that the salient regions of a scene should be its most informative parts. In [52, 53], they introduced a saliency model based on this assumption, which uses Shannon's self-information measure for calculating the visual saliency of image regions. Hou and Zhang [54] proposed an approach based on incremental coding length to measure the entropy gain of each feature. The final saliency map is obtained by maximizing these entropy gain.

According to the widely-used MIT saliency benchmark dataset [12], the current best saliency prediction model without using CNNs is the Boolean Map Saliency (BMS) model, proposed by Zhang and Sclaroff [51]. In their BMS model, an input image is encoded by a set of binary images, i.e. the Boolean maps. The boolean map is generated by randomly thresholding the image's color channels in Lab color space. The surroundedness cue is explored by analyzing the topological structure of Boolean maps. Their underlying idea is that surrounded regions are more likely to be perceived as foregrounds. Compared to other biologically plausible models, BMS is superior in terms of both speed and accuracy.

2.1.1.2 Spectral Analysis-based Saliency Methods

Instead of investigating visual saliency in the spatial domain, models in this category infer saliency in the frequency domain.

In [55], Hou and Zhang introduced a spectral residual saliency model based on the idea that similarities imply redundancies. The spectral residual of an input image is obtained by the difference between its log amplitude spectrum in the Fourier domain and the averaged spectrum, which is approximated by convolving its log spectrum with a local average filter. After Inverse Fourier Transform (IFT), the final saliency map in the spatial domain can be obtained. However, in [56], Guo *et al.* pointed out that the saliency map obtained in the above method can be directly obtained by Phase spectrum of Fourier Transform (PFT) regardless of the amplitude spectrum value. Base on this finding, they introduced a PFT approach to generate the saliency map. In [57], it was argued that these spectral analysis-based methods are equivalent to a local gradient operator followed by Gaussian blurring, and thus cannot detect large salient regions very well. They proposed a saliency method based on spectral scale-space analysis to overcome this limitation.

2.1.1.3 Machine Learning-based Saliency Methods

Most existing saliency models use a set of biologically plausible linear filters for feature extraction. However, these approaches require many manually-designed parameters [58–60]. Therefore, some saliency models with free parameters or non-parametric are proposed by employing machine learning.

In [58], Kienzle *et al.* used a support vector machine (SVM) to learn a saliency model from human eye-tracking data directly. Judd *et al.* [4] proposed a set of low, mid and high-level image features used to define salient locations and use a linear SVM to train a saliency model. In [61], Shen *et al.* described a three-layer of sparse coding network for predicting visual saliency. They claimed that high-level concepts such as faces and texts were learned in the last layer and mid-level features like junctions, textures, and parallelism could be obtained in its second layer.

Deep learning-based methods can be also classified into this category. We notice that the study of saliency prediction has greatly developed with the advent of convolutional neural networks (CNNs). At present, CNN-based saliency models have defeated the classical saliency models in all saliency prediction benchmarks. Moreover, the limitation of the aforementioned machine learning-based approaches is that they are largely dependent on their pre-defined or pre-learned features. In contrast to them, deep learning-based approaches can learn diverse and comprehensive saliency-influential features directly from the larger-scale datasets. Therefore, the deep learning-based saliency methods are presented in an individual subsection.

2.1.2 Deep Learning-based Visual Saliency Prediction Methods

Nowadays, the advances in deep learning have already boosted the progress in saliency prediction. To the best of our knowledge, the first attempt to use CNNs to predict visual saliency was introduced by Vig *et al.* [62]. Their model consists of three individual and different shallow networks (from one layer to three layers) for feature extraction. The final saliency map is obtained by feeding these feature maps into a trained linear SVM. However, this model is inferior to some classic biologically plausible saliency models such as GBVS [50] and BMS [51] models, due

to the limited depth of their networks. After that, researchers seek to use deeper models (e.g. AlexNet [35] in [63, 64], VGGNet [65] in [38, 39], and ResNet [66] in [40, 67]) as the backbone network and utilize the fully convolutional network (FCN) [68] framework for fully leveraging the powerful capabilities of deep CNN (DCNN) models in contextual feature extraction.

The DeepGaze models [63, 69] showed that the deep features from the ImageNet [36] pre-trained VGGNet can be directly used to predict visual saliency without fine-tuning. But other works supported that fine-tuning the pre-trained backbone network can obtain a better performance. In [64], Huang *et al.* proposed a SALICON model that uses a shared backbone network applied at two different image scales for incorporating the multi-scale features. Besides, they used some saliency evaluation metrics as their loss functions. Pan *et al.* [70] proposed a saliency model by using Generative Adversarial Network (GAN). In their model, a generator and a discriminator are working together for producing the saliency map. In [39], Kruthiventi *et al.* introduced an FCN-based saliency model that uses the inception module [1] to capture the multi-scale features. Wang et al. [38] proposed a skip-layer-based saliency model with an encoder-decoder architecture. Three different decoders are built by feeding the features from the top of three different convolutional blocks in the encoder network for multi-level predictions. In [40], Liu *et al.* used a spatial contextual LSTM and two different backbone networks for incorporating both global and scene contexts. Cornia *et al.* [67] proposed to use an attentive convolutional LSTM for iteratively refining the predicted saliency maps. Besides, a linear combination of three different saliency evaluation metrics is adopted as the overall loss function for further improvement.

Currently, DCNN models utilize some down-sampling operations (e.g. max pooling and convolutions with strides) to reduce the computation cost and enlarge the receptive fields of their subsequent layers. For simplification, we denote the ratio of the input image spatial resolution to the output resolution of DCNN by *output_stride*. The more usage of down-sampling operations, the higher *output_stride*. However, DCNN with a higher *output_stride* also means its feature maps in the top layers have a relatively smaller spatial resolution. Such limited spatial information may not support effective dense saliency prediction [40, 67]. A naive approach, presented in [71, 72], to increase the spatial resolution in top layers is directly removing some down-sampling operations in their models. But this simple

approach will unavoidably reduce the receptive field sizes of the subsequent layers. Because the size of the receptive field affects the amount of contextual information for the final saliency inference, such a reduction in receptive field size is suboptimal. Therefore, a trade-off between the spatial resolution of feature maps and the computation cost should be guaranteed while maintaining suitable receptive field sizes. Several powerful deep saliency prediction models [39, 40, 67] adopt dilated convolutions [2, 73, 74] to increase the receptive field sizes of the top layers, while compensating for the reduction in receptive field size induced by removing down-sampling operations.

Previous studies [38, 39] demonstrated that multi-scale contextual features are essential to the visual saliency prediction. The reason for this conclusion is that visual information is processed at various scales by human eyes [1, 37]. Table 2.1 provides a comparison of recent deep saliency models and our proposed model–DINet, which will be presented in Chapter 3. The models with multi-scale inputs will integrate multi-scale contextual features while some models with single input still can capture these due to their adopted multi-scale feature extraction modules, as discussed in the next section.

Most of the existing DCNN-based saliency models directly some the typical pixel-wise classification or regression loss functions. In [75], Jetley *et al.* proposed to use loss functions based on probability distribution (PD) distances with softmax normalization for training saliency models. Their experimental results demonstrated the improvement by considering saliency maps as probability distributions.

Regarding the center-bias phenomenon, some of the deep saliency models learn the center-bias explicitly by their carefully designed modules, such as the location biased convolutional layers in [39]. However, with the help of the large-scale saliency dataset–SALICON [3], DCNN-based saliency models can learn this bias implicitly and solely from the training data [75, 76].

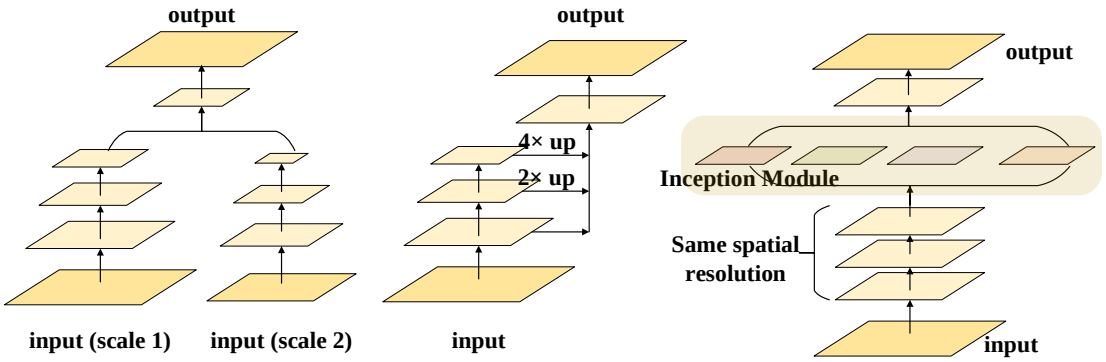
2.1.3 Multi-Scale Feature Extraction Architectures in Visual Saliency Prediction

In Fig. 2.2, we summarize the existing deep architectures aiming at capturing multi-scale contextual features in saliency prediction. These models can be roughly

TABLE 2.1: Overall comparison of recent deep saliency prediction models.

Model	Backbone network	*output_stride*	Input/Inputs	Multi-scale	Loss function	pixel/PD	Center-bias
SALICON [64]	AlexNet/VGG16/GoogleNet	16	multi inputs	yes (IPN)	KLD/CC/NSS	PD	implicit
DeepGazeII [69]	VGG19	16	single input	no	BCE (softmax)	PD	explicit
PDP [75]	VGGNet	N/A	single input	no	PD distances (softmax)	PD	implicit
ML-Net [71]	VGG16	8	single input	yes (Skip-layer)	Euclidean distance	pixel	explicit
SAM [67]	VGG16/ResNet50	8	single input	no	KLD + NSS	PD + pixel	explicit
SALGAN [70]	VGG16	16	single input	no	BCE	pixel	implicit
DeepFix [39]	VGG16	8	single input	yes (Inception)	Euclidean distance	pixel	explicit
DSCLRCN [40]	VGG16/ResNet50 + Places-CNN	8	multi inputs	yes (Skip-layer)	NSS	pixel	implicit
DVA [38]	VGG16	16	single input	yes (Skip-layer)	BCE	pixel	implicit
MxSalNet [72]	VGG16	8	single input	yes (Skip-layer)	Euclidean distance + CCE	pixel	explicit
DINet (Ours) [8]	ResNet50	8	single input	yes (Inception)	PD distances (linear)	PD	implicit

KLD: Kullback-Leibler divergence, PD: probability distribution, BCE: binary cross-entropy, N/A: not available, NSS: normalized scanpath saliency, CCE: categorical cross-entropy.



(a) Image Pyramid Network **(b) Skip-layer Network** **(c) Inception-based Network**

FIGURE 2.2: Illustrations of existing deep learning architectures to capture multi-scale information in saliency prediction.

classified into three categories: i) image pyramid network; ii) skip-layer network; and iii) inception-based network.

2.1.3.1 Image Pyramid Network

The most straightforward way to learn multi-scale feature representations can be found in [64, 77]. Their idea is to apply duplicate or multiple feature extractor networks with the multi-scale inputs, as shown in Fig. 2.2(a). The output features from different network streams are merged and fed into the following decoder network for the saliency inference. Such an image pyramid network (IPN) architecture with multi-scale inputs indeed can capture the multi-scale contextual features. Nevertheless, training and testing these models are not efficient in both computation cost and memory usage due to their multiple feature extractors.

2.1.3.2 Skip-layer Network

Due to the down-sampling operations in the common backbone networks, the features obtained from the different layers of the backbone network will have different spatial resolutions and characteristics due to the different receptive fields [78]. Based on this principle, architectures with skip-layers have been proposed in [38, 63, 71]. Skip-layer network captures multi-scale contextual features by concatenating the outputs of different layers with increasingly larger receptive fields and *output_stride*, as illustrated in Fig. 2.2(b). More importantly, the skip-layer network can efficiently utilize intermediate features while the conventional approach

only utilizes the top-most features. Despite the high efficiency, the main problem in the skip-layer network is that spatial information gradually reduced in the higher layers due to the usage of down-sampling operations. Direct up-sampling and concatenating these feature maps from different layers without feature adaptation will bring uncertainty and ambiguity into the saliency inference, consequently leading to a limited performance gain.

2.1.3.3 Inception-based Network

As demonstrated in Fig. 2.2(c), inception-based network [39] avoid the above problem by utilizing the dilated convolutions and removing some down-sampling operations in the backbone network. Therefore, the top-most features still have sufficient spatial information to support the dense prediction. Inception modules, proposed in the well-known GoogleNet [1], are attached to the top of the backbone network to capture multi-scale contextual features. The main idea of the inception module is to use convolutions with multiple kernel sizes. As such, the local contextual information in different receptive fields can be extracted. In general, using the inception module can help the saliency model to obtain larger performance gain than the skip-layer network and much smaller computation cost than the IPN-based models. However, the original inception module is still not very economic in both computation and optimization.

Our work, which will be presented in Chapter 3, is based on this type of network. In particular, we revise the original inception module to have a more powerful multi-scale feature extraction capacity in a computationally-friendly manner, as will be discussed in Section 3.2.3. In addition to the GoogleNet [1], our dilated inception module also takes the advantage of the atrous spatial pyramid pooling (ASPP) module in the DeepLab model [2] for semantic segmentation. We apply those parallel dilated convolutional layers to form our dilated inception module and thus obtain state-of-the-art saliency prediction performance in terms of both speed and accuracy.

2.2 Salient Object Detection

Salient object detection (SOD) aims to segment the entire salient foreground objects from the background [31]. The saliency maps generated by SOD models and the results produced by fixation prediction models are quite different. In this section, we first briefly review the classic SOD methods. Then, we focus on investigating the CNN-based SOD models. In particular, those deep models with feature aggregation modules are presented. Finally, the loss functions in CNN-based SOD models are discussed. The main differences between our work in Chapter 4 and other existing works are shown in this subsection.

2.2.1 Classic SOD Models

The study of saliency detection also starts with Itti's model [49]. At this early stage, the saliency models for fixation prediction are interchangeably used in saliency detection. Until [30], Liu *et al.* firstly formulated saliency detection or SOD as an image segmentation problem. Research has shown that SOD, which emphasizes the object-level integrity of saliency detection results, is more useful and suitable for a wide range of object-level computer vision applications. After that, numerous saliency models, especially for SOD, are emerged in the past two decades.

Classic SOD models mainly rely on hand-crafted features and heuristic cues to detect salient objects. Some typical works are as follows. In Liu's work [30], they proposed to use a set of low-level hand-crafted features including multi-scale contrast, center-surround histogram, and color spatial distribution to describe a salient object locally, regionally, and globally. A Conditional Random Field (CRF) was applied to combine these features effectively for SOD. Goferman *et al.* [79] used a patch-based approach to incorporate the global context for detecting the salient regions. In [80], Wang *et al.* estimated local saliency by leveraging a dictionary learned from other images, and global saliency by using a dictionary learned from other patches of the same image. Cheng *et al.* [81] proposed a region contrast-based method which measures global contrast differences and spatial coherence simultaneously. In [82], Jiang *et al.* utilized the difference between the regional

color histogram and object-level shape prior to measuring saliency. Later, they further proposed a multi-scale regional features-based approach in [83] for integrating three different regional features together to obtain the final saliency map.

Besides, several extrinsic cues can be utilized in saliency models as supporting information to assist the detection of salient objects. As a result, these extrinsic cues, such as depth cues, temporal relationships, or inter-image correspondence, are available to extend image saliency detection to RGBD saliency detection, video saliency detection, or co-saliency detection. A detailed survey of these extended saliency detection tasks is discussed in [84].

2.2.2 CNN-based SOD Models

Unlike the majority of classic SOD methods based on some heuristic cues, CNN-based approaches are recently adopted by many researchers due to their powerful automatic feature extraction capability and end-to-end characteristic. Moreover, the advances in deep learning techniques have substantially boosted the progress in CNN-based SOD models. According to a recent survey [85], the CNN-based SOD models can be roughly classified into two main categories: Multi-layer Perceptron (MLP)-based methods and Fully Convolutional Network (FCN)-based methods. Some typical methods of these two categories are as follows.

2.2.2.1 MLP-based Methods

Early CNN-based SOD models attempt to search for salient objects by extracting features from the local image patches or superpixels and performing the saliency inference by an MLP-classifier.

In [86], Li *et al.* proposed to extract multi-scale CNN feature vectors from three nested windows with an ImageNet pre-trained deep neural network. An MLP is trained on the top to regress these concatenated multi-scale CNN features into the region-level saliency. In their model, an input image is decomposed into a set of non-overlapping regions. The final saliency map of this input image is fused by multiple region-level saliency maps from the MLP classifier. Zhao *et al.* [87] used a deep model with two branches to extract local and global context separately from two superpixel-centered windows of different sizes. By using two individual MLPs,

it can output local and global saliency detection results. The final saliency maps are obtained by fusing these two results via a fully-connected layer in the local branch. In [88], He *et al.* extracted two hand-crafted feature sequences, including color uniqueness sequence and color distribution sequence, as the input features for each multi-scale superpixel. These hand-crafted features are further processed by two different CNNs and inferred together by multiplication to produce the binary saliency scores. The final saliency map is obtained by weighted summing these inferred saliency scores of all the scales.

However, due to the usage of MLP-classifier, these MLP-based methods suffer from the loss of spatial information, which unavoidably results in some coarse and imprecise saliency maps. Moreover, they are also time-consuming as they need to process all of the local image regions one by one.

2.2.2.2 FCN-based methods

Lately, with the advent of the fully convolutional network (FCN) [68], the latest SOD models adopt this FCN framework and directly process the whole input image to overcome the aforementioned drawbacks in MLP-based methods. A summary of the recent FCN-based SOD models is listed in Table 2.2. Here, we first present some typical works in this table. Then, those FCN-based models with feature aggregation modules, which are mostly related to our work in Chapter 4, are discussed in the latter part.

In [89], Liu *et al.* used an FCN-based backbone network to get a coarse global prediction for each image. These coarse predictions, which are integrated with local context information, are hierarchically and progressively refined by using recurrent convolutional layers. Li *et al.* [91] proposed to combine a pixel-wise prediction from a multi-scale FCN sub-network and a superpixel-wise prediction from a modified MLP-based sub-network with the shared feature extractor [86]. In [95], Tang *et al.* proposed a similar approach to combine pixel-level saliency prediction and superpixel-level saliency estimation with multiple CNNs. Liu *et al.* [104] proposed to hierarchically embed global and local pixel-wise contextual attention modules in a U-Net [108] framework.

Besides, some of the methods adopt the multi-task learning framework into SOD for further improvement. In [92], Kruthiventi *et al.* proposed a unified FCN-based

TABLE 2.2: Overall comparison of recent FCN-based SOD models.

Method	Year	Backbone	Architecture	Feature Aggregation	Multi-task	Loss function	Training Dataset	#Training	CRF
DHSNet [89]	2016	VGGNet	FCN + Recurrent convolutional layers	skip-layers	no	BCE	MSRA10K [81] + DUT-OMRON [90]	6,000+3,500	
DCL [91]	2016	VGG-16	MS-FCN + MLP-based saliency model [86]	skip-layers + dilated convolutions	no	BCE	MSRA-B [30]	2,500	✓
SU [92]	2016	VGG-16	FCN + Inception modules	skip-layers + inception modules	Yes (eye fixation prediction)	Not Metioned	MSRA10K [81] + SALICON [3]	10,000+15,000	✓
RFCN [93]	2016	VGGNet	Recurrent FCN	skip-layers	Yes (pretrained on semantic segmetation)	BCE	PASCAL VOC 2010 [94] + MSRA10K [81]	10,103+10,000	
CRPSD [95]	2016	VGGNet	FCN + MLP-based saliency model [87]	skip-layers	no	BCE	MSRA10K [81]	10,000	
UCF [96]	2017	VGGNet	FCN Encoder + FCN Decoder	no	no	BCE	MSRA10K [81]	10,000	
Amulet [97]	2017	VGG-16	FPN + RFCs + Boundary preserved refinement	RFCs	no	Multiple BCE	MSRA10K [81]	10,000	
NLDF [98]	2017	VGG-16	FCN + Contrast feature extraction + deconvolution layers	multi-resolution grid structure	no	BCE + boundary IoU loss	MSRA-B [30]	2,500	✓
DSS [99]	2017	VGG-16	FCN + Skip connections	skip-layers	no	Multiple BCE	MSRA-B [30]	2,500	✓
SRM [100]	2017	ResNet-50	2 FCNs + Pyramid pooling module	pyramid pooling module	no	Multiple BCE	DUTS [101]	10,553	
BDMP [102]	2018	VGG-16	FPN + Gated bi-directional message passing modules	dilated convolutions	no	BCE	DUTS [101]	10,553	
DGRL [103]	2018	ResNet-50	FCN-based recurrent localization network +Boundary refinement	skip-layers + inception modules	no	BCE	DUTS [101]	10,553	
PiCANet [104]	2018	VGG-16/ResNet-50	U-Net + Local and global pixel-wise attention modes by bidirectional LSTMs	skip-layers + attention	no	Multiple BCE	DUTS [101]	10,553	✓
PAGRN [105]	2018	VGG-19	FCN + Multi-path attention guided recurrent network	skip-layers + attention	no	Not Metioned	DUTS [101]	10,553	
ASNet [32]	2018	VGG-16	FCN + a hierarchy of convolutional LSTMs	skip-layers	Yes (eye fixation prediction)	Multiple (KLD + (BCE +F-measure losses))	SALICON [3] + MSRA10K [81] + DUT-OMRON [90]	15,000+10,000+5,168	
PAGENet [85]	2019	ResNet-50	FPN + Pyramid attention module + Salient edge detector	skip-layers + attention	Yes (salient edge detection)	Multiple (L2 norm + weighted BCE))	MSRA10K [81]	10,000	✓
BASNet [43]	2019	ResNet-34	U-Net + Residual refine module	skip-layers	no	Multiple (BCE + SSIM + IoU loss)	DUTS [101]	10,553	
CPD [106]	2019	VGG-16/ResNet-50	FCN Encoder + 2 FCN Decoders + Holistic attention module	skip-layers + attention	no	Multiple BCE	DUTS [101]	10,553	
PoolNet [5]	2019	VGG-16/ResNet-50	FPN + Pyramid pooling modules + Feature aggregation modules	skip-layers + inception modules	Yes (salient edge detection)	BCE + balanced BCE	DUTS [101] + HED-BSDS_PASCAL [5]	10,553	
EGNet [6]	2019	VGG-16/ResNet-50	FPN + Salient edge features extraction module	skip-layers	Yes (salient edge detection)	Multiple BCE	DUTS [101]	10,553 + 42,346	
Ours [107]	2020	ResNet-50	FPN + MS-FAMs + FCN Decoder	skip-layers + MS-FAMs (DIM+BAM)	no	(BCE + Dice) + PG loss	DUTS [101]	10,553	

BCE: binary cross-entropy, FCN: fully convolutional network, FPN: feature pyramid network, RFC: resolution-based feature combination module, CRF: conditional random field,.

model for predicting eye fixations and segmenting salient objects. The first branch learns to infer visual saliency from the top-most features, while the SOD branch aggregates multi-level features to detect salient objects. Wang *et al.* [32] followed this idea and utilized a hierarchy of convolutional LSTMs to iteratively infer the salient object segmentation. More importantly, the learned fixation map is used for guiding accurate object-level saliency estimation in a top-down way. In [85], they further proposed to use salient edge detection, instead of fixation prediction, as the auxiliary task for the SOD. The edge detection module can provide explicit edge information, which can be used to locate salient objects and sharpen their boundaries.

The existing works which are most relevant to our work are the FCN-based models with feature aggregation modules. At present, many works [106, 108, 109] have shown that aggregating multi-level and multi-scale features into the saliency inference can further improve the performance of their methods. As such, various feature aggregation approaches to achieve this goal have emerged.

In [100], Wang *et al.* proposed a multi-stage saliency model for progressively refining the coarser saliency maps obtained at the early stages. The pyramid pooling module is adopted to exploit global context information for feature aggregation. Zhang *et al.* [97] aggregated multi-level convolutional features into multiple resolutions by their proposed resolution-based feature combination modules for simultaneously incorporating coarse semantics and fine details. These multiple aggregated features are further fused and refined in a top-down manner with deep supervision. In [99], Hou *et al.* introduced a series of short connections to their skip-layer architectures for fusing the multi-level features extracted from FCN. Luo *et al.* [98] proposed to use a multi-resolution grid structure to combine local contrast and global information. In [105], Zhang *et al.* proposed an FCN-based saliency model with multi-path recurrent connections and two attention mechanisms for selectively integrating contextual information from multi-level features to generate powerful attentive features. Liu *et al.* [5] designed a pooling-based feature aggregation module to fuse and refine the multi-level features in a top-down manner. We observe that most of them directly fused features with different levels by simply using upsampling followed by sum or concatenation operations. However, as pointed by [106], some low-level features may contribute less to the performance of

feature aggregation methods. Such directly fusing may degrade the discrimination ability of the aggregated features.

To address the above problem, we propose to use a branch-wise attention mechanism (BAM) combined with the modified dilated inception module (DIM) [8] to form our multi-scale feature aggregation module (MS-FAM) [107] for highlighting the discriminative features and suppressing those features which may confuse the later saliency inference in an adaptive manner.

2.2.3 Loss Functions in SOD

Most SOD methods use binary cross-entropy (BCE) as their training loss. But BCE loss is a typical pixel-wise loss function which only accounts for the pixel-wise difference between labels and predictions. As a result, it does not consider the spatial relationship of label distribution and equally weights both the foreground and background pixels. There are two main drawbacks in training SOD models with BCE loss: Firstly, the foreground pixels are accumulated within the salient objects which have some clear boundaries away from the background. BCE cannot help SOD models to uncover this relationship and hence leads to blurry boundaries and some miss-detected regions within the complete salient objects. Secondly, SOD is a class-imbalanced task, as evidenced by the fact that the number of salient pixels is much smaller than the non-salient ones in a labeled saliency map. Models trained with BCE or other similar pixel-wise losses would have biased prior due to the biased label distribution and tend to predict unknown pixels as the background, consequently leading to some incomplete predictions. Some of the recent works [6, 99, 106] applied the deep supervision by utilizing the ground truth saliency maps to guide the intermediate predictions with multiple BCE Losses. But the performance gain of this technique is not obvious as there is lacking guidance towards characterizing the spatial dependencies.

There are several attempts to alleviate these drawbacks. One possible way is to seek some more suitable losses in training SOD. In [41], Zhao *et al.* proposed to directly maximize the F-measure for SOD. Since F-measure is a widely adopted evaluation metric in SOD, models trained with their F-measure loss can achieve better performance and easily adjust the compromise between precision and recall by changing the β^2 factor in this loss. Qin *et al.* [43] proposed a hybrid loss,

which is fused by the BCE, SSIM, and IoU losses. Equipped with this hybrid loss, their SOD models can be able to capture both large-scale and fine structures. However, most of these alternative losses are not specifically designed for capturing the structure difference and modeling spatial dependency.

Another way is to introduce multi-task learning loss into SOD, as presented in the last section. Recent works [5, 43, 85] usually take the edge detection task, instead of fixation prediction, as an auxiliary task for the SOD. They have shown that the edge information can be leveraged for locating salient objects and sharpening their boundaries. However, these methods need to build an additional sub-network for predicting fixations or detecting the edges which unavoidably increase their inference time for detecting salient objects.

Our proposed progressive self-guided loss [107] is quite different from the above approaches. Instead of exploring new losses or introducing multi-task learning techniques, we address the above-mentioned limitations from a novel perspective by providing a series of new progressive and auxiliary training supervisions. These newly training targets are generated from the network predictions but with slightly better shapes. More importantly, they are not fixed and progressively optimized in a region growing manner for guiding the SOD models to uncover the spatial dependencies. As such, the SOD model trained with our PSG loss can be progressively guided to highlight the entire salient objects without architecture modification.

2.3 Image Quality Assessment

Image quality assessment (IQA) aims to evaluate the perceptual quality of a digital image in a manner that is consistent with human subjective opinions. According to the accessibility of the pristine reference images, IQA models can be classified into full-reference (FR) [110–113], reduced-reference (RR) [114–116], and no-reference (NR) [18, 117–120] three types. Among them, NR-IQA has a broad range of application scenarios since reference images are not accessible in most practical applications. In this section, we review some typical NR-IQA methods and discuss the relationship between visual saliency and NR-IQA. Our work in Chapter 5 is the first attempt to put saliency prediction and quality evaluation together via a multi-task learning framework.

2.3.1 NR-IQA Methods

According to the type of extracted features, NR-IQA methods can be roughly classified into three groups: Natural Scene Statistics (NSS)-based methods, feature learning-based methods, and CNN-based methods. NSS-based approaches are based on the assumption that distortion-free natural images have inherent statistical regularities and the presence of distortions in natural images will change such regularities. Complex statistical models for wavelet coefficients [121] or discrete cosine transform coefficients [122] or locally normalized luminance coefficients [123] were developed for NSS modeling to extract quality-aware features. However, the NSS-based NR-IQA methods heavily depend on the domain knowledge on NSS modeling which is too complex to achieve a sufficient understanding.

Different from those NSS-based NR-IQA methods, the quality-aware features are automatically learned from data in the feature learning-based NR-IQA methods. According to the utilization of quality information in the feature learning process, the existing feature learning-based NR-IQA methods can be classified into unsupervised and supervised feature learning schemes. In CORNIA [124] and HOSA [118], a codebook pre-learned in an unsupervised manner (i.e., K-means clustering) was used for feature encoding to generate quality-aware features directly from local normalized image patches. In MSDD [18], multi-stage discriminative dictionaries pre-learned in a quality supervised manner were used for feature encoding to generate quality-aware features. Experimental results have demonstrated the unique advantages of the feature learning-based NR-IQA methods as compared to the previous approaches using hand-crafted NSS features. However, with the advent of deep CNN, there's no doubt that these feature learning-based methods were surpassed by the deep CNN-based solutions for NR-IQA.

Kang et al. [46] firstly implemented a shallow CNN model to extract features on small image patches for NR-IQA. The final quality score of an input image was computed by averaging the predictions of all patches cropped from it. However, this method and its successors suffer from the label noise problem caused by assigning the global quality label for all the patches cropped from the same input image. To partially represent the region-wise perceptual quality variation, Kim and Lee [125] proposed to use one of the classical FR-IQA methods to generate local quality scores for image patches as the local ground truth targets. However, such a strategy

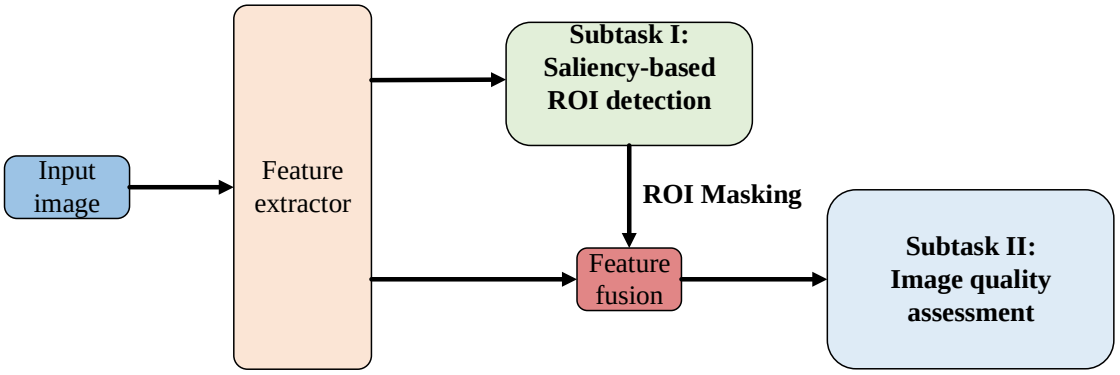


FIGURE 2.3: An illustration of our multi-task learning framework.

is not applicable to the real-world authentically distorted images which usually do not have the corresponding reference versions.

To learn a better feature representation, recent CNN-based NR-IQA methods proposed to adopt the multi-task learning framework. Kang et al. [44] further extended their work [46] to estimate image quality and distortion type simultaneously via a classic multi-task CNN. Although these two sub-tasks are jointly optimized, there is no interaction between these two sub-tasks in their designed network. To address this problem, Ma et al. [45] proposed a new multi-task deep CNN model for NR-IQA. In their model, the quality evaluation sub-task depends on the outputs of the distortion identification sub-task. As such, the distortion type information is transparent to the primary quality evaluation sub-network for better quality evaluation. Our model also follows this pipeline where two sub-tasks are jointly optimized and have certain dependencies on each other. The main difference between ours and Ma's method is that we adopt saliency prediction to replace their distortion identification as the complementary sub-task for providing more universal yet closely related perceptual information to support the primary quality evaluation sub-task. More importantly, the inability of existing multi-task CNN-based NR-IQA methods on evaluating the qualities of authentically distorted images is overcome by using our approach. A simple illustration of our proposed multi-task learning framework is depicted in Fig. 2.3.

2.3.2 Visual Saliency for NR-IQA

It is known that visual attention for different regions in an image is non-uniform and people tend to focus on visually salient areas while assessing image quality

[17, 113, 126, 127]. The relevance between visual saliency and IQA has been validated by some subjective psychophysical studies [128–130]. Alers et al. [128] conducted a series of eye-tracking and quality evaluation experiments to explore their relationship. The subjects in this study were firstly required to label the qualities of the images with different levels of distortion. With the help of the eye-tracker, the eye-tracking data of these subjects were also recorded during the labeling stage. The regions in these images can be classified as the regions of interest (ROI) or background according to the eye-tracking data. After that, a new set of images with the same original content, but with a different level of quality for the ROI and background are created for the next round of labeling. Their results showed that the quality of the manipulated images mainly dependent on the quality of their ROI. It can demonstrate that the visual distortions that happened in salient regions are more visually unpleasant than those in non-salient regions. In [129, 130], Zhang et al. found that the occurrence of distortion in an image tends to deviate fixation deployment and the extent of distortion determines the amount of saliency deviation. As a result, visual saliency is also affected by visual distortions.

Due to the above relevance, the local visual importance measure has been considered in some NR-IQA methods to capture such spatial attention variations for better quality evaluation [127, 131–133]. Liu et al. [131] used the ground truth eye-tracking data as the weighting information in classic IQA metrics. A large amount of performance gain was achieved by adding such additional saliency information in them. Zhang et al. [132] utilized the classical visual saliency models to obtain the local weights (i.e., saliency map) of an image. These local weights were integrated into one of the existing traditional IQA metrics to assess the quality of this image via the obtained weighted local quality map. By selecting an appropriate saliency model, these saliency-guided IQA metrics can outperform their original version significantly. However, even the saliency-guided versions of these classic NR-IQA methods were still not comparable with those aforementioned CNN-based approaches.

Most of the CNN-based NR-IQA methods, which are trained on local image patches, simply assign the subjective quality score of an image to all the local patches cropped from it as their local quality label for training their networks. Although

this patch-based training strategy can provide enough training samples for training a deep NR-IQA model compared to the image-based one, it is still problematic because the local perceptual quality is not well-defined and not always consistent with the global quality score [45]. For this consideration, VIDGIQA [127] and WaDIQaM-NR [133] are proposed to train the deep IQA models by jointly learning the local visual importance and quality score of each local patch. The learned local importances are used as the weights for the quality scores of local patches to estimate the final global quality by weighted averaging. However, these two local immediate regression targets are jointly optimized only with the single global quality scores as supervision. We argue that their newly introduced local visual importance weights are still not well-defined, as evidenced by there are no direct supervisions for training these local immediate components.

Unlike the above-discussed methods, our SGDNet model is an image-based approach that generates feature maps from the whole input images instead of their local patches to avoid the potential problem of label noises. More importantly, we utilize the proxy saliency maps produced by a teacher saliency model–our DINet [8], which are trained on the large-scale saliency prediction dataset–SALICON [3], to serve as the direct supervisions. Therefore, the shortage of training data is partially alleviated by introducing such more informative labels in the multi-task learning framework.

Chapter 3

Dilated Inception Network for Visual Saliency Prediction

3.1 Introduction

Visual saliency prediction[1], also known as eye fixation prediction, aims to compute a saliency map that topographically represents humans' attentional priority when they view a given image [48]. Predicting human eye fixations is important to understand and simulate the behavior of visual attention for advancing a wide range of visual-oriented multimedia applications such as image retrieval [134], image retargeting [14], video summarization [20], image and video compression [15, 16], visual quality assessment [17–19], virtual reality content design [21], and more.

Most of classic saliency prediction models [49–51] are biologically plausible. They mainly adopt multiple low-level hand-crafted features, such as intensity, color, and so on, and combine these features in a heuristic way (e.g. multi-scale center-surround contrast [49], graph-based random walk [50], etc.). However, these low-level hand-crafted features and their heuristic combinations are insufficient to represent the wide variety of factors that contribute to visual saliency [39, 40, 67].

With the advent of Deep Convolutional Neural Network (DCNN), the feature extraction and combination pipeline could be simplified in an end-to-end manner and a data-driven automatic feature learning is also available. At present, DCNN-based

[1]The work in this chapter has been published in [8].

saliency models have outperformed the classical saliency prediction models in all challenging saliency prediction benchmark datasets [3, 4, 10]. Within these DCNN-based models, the use of multi-scale contextual features [38–40, 72], which aims to characterize the diverse saliency-influential factors at different receptive field sizes, make them outstanding. However, these deep saliency models suffer from the huge computation cost problem caused by fully exploiting the comprehensive feature representations.

In this work, we propose a DCNN architecture called Dilated Inception Network (DINet) for visual saliency prediction. To fully exploit the multi-scale contextual features, an efficient yet effective dilated inception module (DIM) is implemented. The original inception module [1] utilizes multiple convolutional layers with different kernel sizes to serve as multi-scale feature extractors with various receptive fields. In contrast, our DIM uses parallel dilated convolutions with different dilation rates [2] to capture more comprehensive and effective multi-scale contextual features with much less computation cost. Our DINet is built by an encoder-decoder framework where the encoder consists of the DCNN-based backbone network and our DIM and the decoder network for final saliency inference is a simple fully convolutional network.

A recent study [75] shows that training a DCNN-based saliency model with their proposed softmax normalization-based probability distribution (PD) distance metrics results in superior performance compared to using commonly-used pixel-wise regression loss functions. Instead, we further propose a set of linear normalization-based PD distance metrics as the new learning objectives to outperform both of them. An extra performance gain is achieved by providing an additional linear regularization to formulate the saliency prediction task as a PD prediction problem.

The performance of our DINet is evaluated on various saliency prediction benchmark datasets. The peer comparison results indicate that our DINet can achieve state-of-the-art performance in terms of both accuracy and speed. The source codes of DINet and its pre-trained model are publicly available[2].

In summary, the contributions of this work are threefold:

- We propose an efficient and effective dilated inception module (DIM) to capture the multi-scale contextual features. The scale diversity is enriched by

[2]https://github.com/ysyscool/DINet

introducing paralleled dilated convolutions with various dilation ratios at lower computation cost. In particular, the effectiveness of this DIM can be further verified by our proposed visualization method and model ablation analysis.

- A set of linear normalization-based probability distribution distance metrics are proposed as loss functions to optimize our DINet. An additional linear regularization is introduced by them, consequently leading to a promising performance gain.
- The computation cost is further reduced by replacing the deconvolutional layers with a fully convolutional decoder network. As a result, the whole model is efficient to achieve real-time performance.

3.2 Dilated Inception Network

In this section, we present the architecture of our DCNN-based saliency prediction model–DINet (Dilated Inception Network). The whole model is depicted in Fig.3.1. Our model starts from the Dilated Residual Network (DRN) [74] which is the primary feature extractor to extract dense feature maps with relatively larger spatial resolution. Our proposed dilated inception module is attached to the top of the DRN for capturing the multi-scale features. A simple yet effective decoder network is employed at the end for saliency inference. Furthermore, since the saliency map can be viewed as a probability distribution (PD) of human fixations, we propose a set of linear normalization-based PD distance metrics for training our DINet to better measure the gaps between our saliency predictions and ground truths.

3.2.1 Dilated Convolution and Dilated Residual Network

3.2.1.1 Dilated Convolution

The main idea of dilated convolution is to insert holes (zeros) in the normal convolutional kernels to increase its receptive field with the same computation and memory cost. Since dilated convolutions are widely used in our model as the core technique, we simply revisit its concept and properties here.

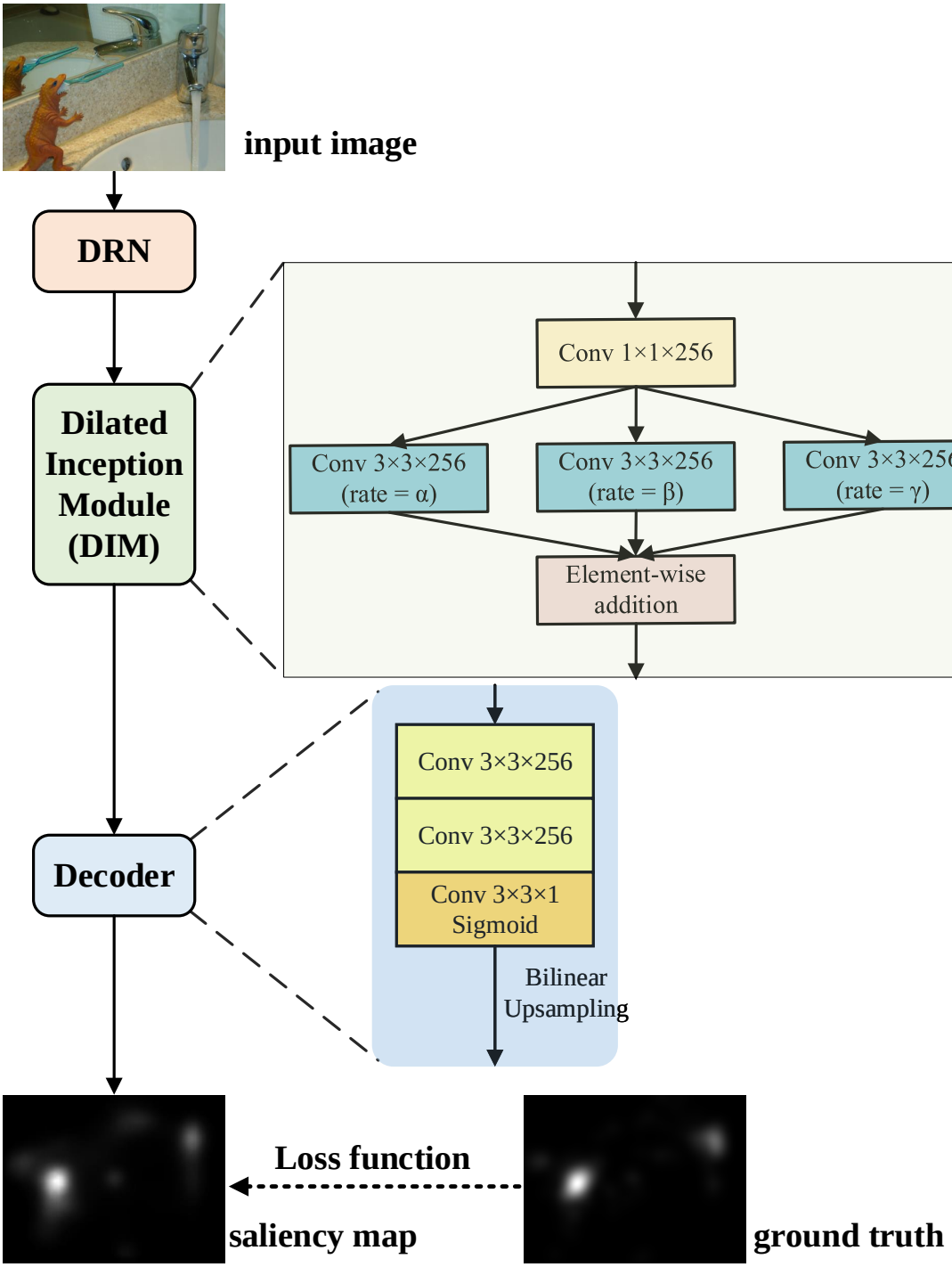


FIGURE 3.1: Architecture of our proposed DINet model for visual saliency prediction.

In general, for each spatial location i, dilated convolution is defined as:

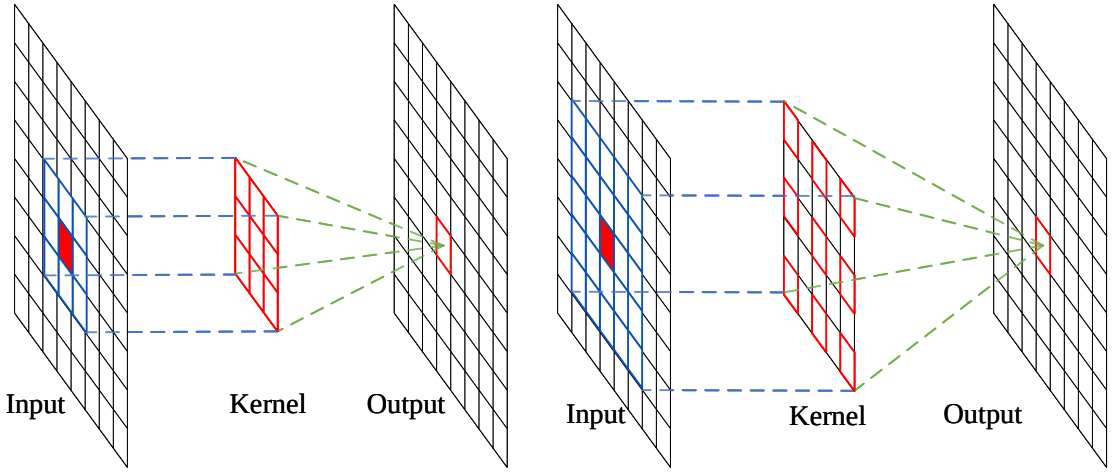
$$y[i] = \sum_{l} x[i + r \cdot l]w[l], \tag{3.1}$$

where $y[i]$ and $x[i]$ denote the output and input on location i, respectively. w is the convolutional filter and r is the dilation rate to sample the input. Dilated convolution is implemented by inserting $r - 1$ zeros between two consecutive spatial positions in the original filter w along each spatial dimension. For a $k \times k$ convolutional kernel, the actual kernel size of a dilated convolution is $k_d \times k_d$, where $k_d = k + (k - 1) \cdot (r - 1)$. It should be noted that dilated convolution still only has $k \times k$ non-zero kernel parameters. As a result, the standard convolution can be viewed as a special case of dilated convolution with $r = 1$. A visual comparison between the standard convolution and dilated convolution is illustrated in Fig. 3.2. A dilated 3×3 convolutional kernel with $r = 2$ sample the feature maps like a 5×5 standard convolutional kernel, which means the receptive fields of the outputs after these two kernels are roughly the same. With this observation, we can arbitrarily change the field-of-view (FOV) of dilated convolutional kernels via choosing different dilation rates under the same number of parameters. By incorporating dilated convolutions into the encoder network, the dilated encoder network is capable of preserving the spatial resolution and compensate for the reduction of receptive field caused by removing some down-sampling operations in the original encoder network.

3.2.1.2 Dilated Residual Network

VGG-16 and ResNet-50 are two commonly used backbone networks for saliency prediction. Besides, both of these two backbone networks have the corresponding dilated versions. Thanks to the residual learning introduced by He et al. in [66], ResNet can be trained very deeply for more comprehensive feature extraction. Existing works also support that (dilated/plain) ResNet-50 based saliency models perform better than those based on (dilated/plain) VGG-16. In this work, we directly employ ResNet-50 as our backbone network.

ResNet-50 network has five blocks of convolutional layers. The *output_stride* of the plain ResNet-50 is 32 which will lead to some ambiguities in dense predictions due to the huge loss of spatial information. In the dilated ResNet-50 [40, 67], to obtain



(a) Standard 3×3 Convolution (b) Dilated 3×3 Convolution (rate = 2)

FIGURE 3.2: A visual comparison between standard convolution (a) and dilated convolution (b). The blue regions in the inputs can be viewed as the receptive fields of the pixels in the outputs.

a relatively larger spatial resolution with an affordable computation cost increase, the configuration of its first three convolutional blocks are kept fixed while the last two blocks, Conv4 and Conv5, are modified by removing the down-sampling operations and replacing some of the standard convolutions inside these blocks by dilated convolutions with dilation rate of 2 and 4, respectively. As a result, the *output_stride* of dilated ResNet-50 is 8 which results in a good compromise between the spatial resolution and computation cost.

3.2.2 Decoder Network

In our encoder-decoder framework, the role of the above-mentioned dilated residual network (DRN) is a basic encoder network. To perform the complete saliency inference, a decoder network is needed to generate the saliency map from the encoded features in this DRN. The conventional decoder network is built by stacking deconvolutional layers which can also help in up-sampling the coarse feature maps into dense ones. However, up-sampling these coarse feature maps by deconvolutions inevitably need longer inference time and also bring some non-smoothing patterns inside them [135]. Thanks to the DRN, the encoded feature maps in our framework have relatively denser spatial information. As a result, the deconvolutional layers are not adopted in our decoder network.

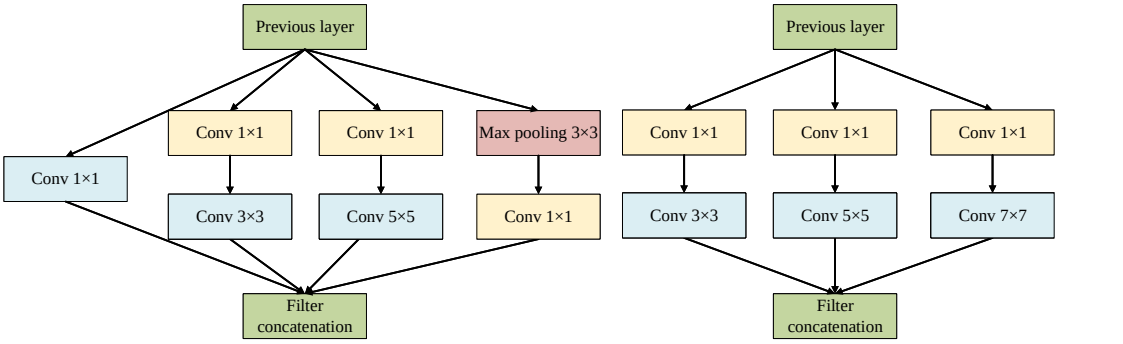
Instead, our designed decoder network is much simpler since it only consists of three stacked standard convolutional layers with one bilinear up-sampling operation in the end. The number of convolutional layers is determined by our experiments in Section 3.3.6.2. In our decoder, the first two layers have 256 3×3 convolutional kernels with the ReLU activation. The last convolutional layer is the prediction layer, which has only one 3×3 convolutional kernel with the sigmoid activation to generate the down-sampled version of the saliency map. The sigmoid activation function is used to rescale the outputs of our saliency model into the target value range. A bilinear up-sampling operation is applied at the end of our model to generate saliency maps with their original resolutions.

The baseline model for this work is the combination of the DRN and this decoder network. To further reduce the number of parameters in this baseline model, a $2048 \times 1 \times 1 \times 256$ convolutional layer is inserted between them. To our surprise, the performance of our baseline model has no visible change with such modification. For constructing our DINet, we replace this newly inserted layer with the proposed dilated inception module, as presented in the next section.

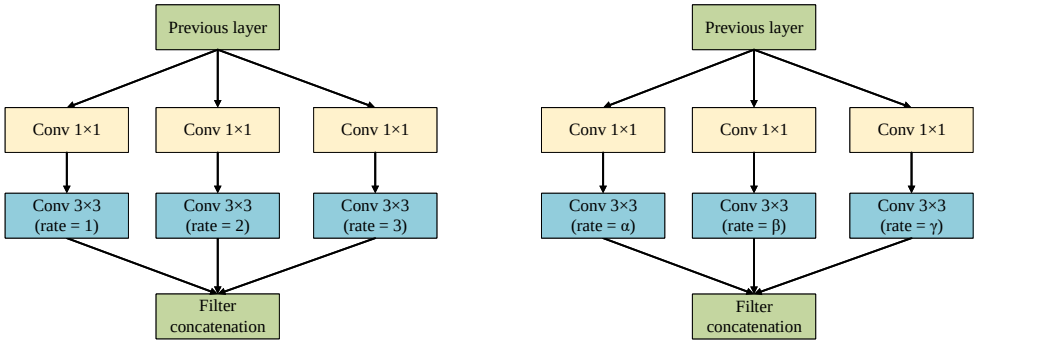
3.2.3 Dilated Inception Module

The proposed module is derived from the inception module which intends to capture the multi-scale contextual information from the inputs [1]. The principal idea of the original inception module is to utilize multiple convolutional layers with different kernel sizes for working as a multi-scale feature extractor to extract contextual features at various receptive field sizes, as shown in Fig. 3.3(a). Unlike the well-known GoogLeNet [1] which is built by stacking several customized inception modules with carefully designed topologies, the proposed inception module works as a single plug-in module in our work to diversify the receptive fields of those encoded features from the output of DRN.

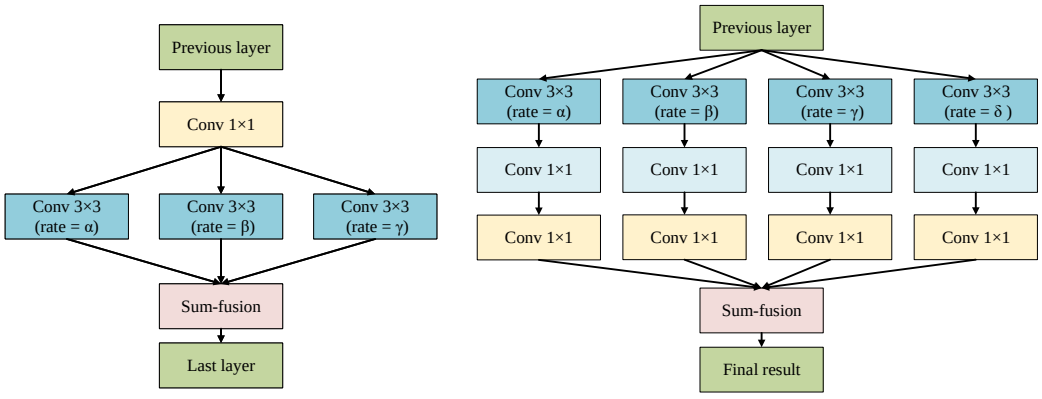
For simplification, the filter numbers in our inception modules are all fixed to 256. By inserting the original inception module between the DRN and decoder network, the performance of our new model is improved obviously with acceptable extra parameters and computations. However, we find that the branch of 1×1 convolutional block has a limited influence on final results. Besides, we replace the max-pooling branch by one 7×7 convolutional layer after one 1×1 convolutional



(a) Original Inception Module **(b) Modified Inception Module**



(c) Basic Dilated Inception Module **(d) Dilated Inception Module-LargerFOV**



(e) Proposed Dilated Inception Module **(f) DeepLab-ASPP Module**

FIGURE 3.3: Inception module and its variations. Module (a) is the original inception module [1]. Modules (b), (c), and (d) are three variants. Module (e) is our final proposed dilated inception module (DIM). Module (f) is the DeepLab-ASPP module [2]. The yellow 1×1 convolutional blocks have the ability of dimensionality reduction.

layer to only investigate the convolutional layers within the inception module, as shown in Fig. 3.3(b). With the help of the 7×7 convolutional block, the modified inception module can extract more diverse and wider field-of-view (FOV) features. To compare the computation cost intuitively, the number of the parameters of a $256 \times 1 \times 1 \times 256$ convolutional layer (without bias term) is denoted as W. Therefore, 7×7 convolutional layer in inception module (b) has $7^2W = 49W$ parameters to be determined, which is much larger than 5×5 convolution with $25W$ parameters and 3×3 convolution with only $9W$ parameters. The total number of parameters in the modified inception model needs an additional $32W$ parameter compared to the original inception model, which results in larger computation cost and longer inference time.

Dilated convolutions, as introduced in Section 3.2.1, can be used to replace the standard convolutions with large kernel sizes under the same receptive field, as shown in Fig. 3.3(c). 7×7 and 5×5 convolutions in the modified inception module can be replaced by 3×3 dilated convolutions with a dilation rate of 3 and 2, respectively. By adopting this replacement, this basic dilated inception module (DIM) can perform similar or even better results as the modified inception module with $(7^2 + 5^2 - 2 \times 3^2)W = 56W$ parameters less. It is worth noting that dilated convolutions in DRN are used in a cascaded way to preserve the spatial resolution and compensate for the reduction in receptive fields. While in DIM, dilated convolutions are used in a parallel way to enhance the encoded features with diverse and comprehensive FOVs.

Besides, the dilation rates of these three parallel dilated convolutions can be arbitrarily changed, as denoted by $[\alpha, \beta, \gamma]$. In our experiments, we empirically set $[\alpha, \beta, \gamma] = [4, 8, 16]$ which has an obvious improvement from the basic dilated or original inception module. This DIM with larger FOV is depicted in Fig. 3.3(d). We further reduce the computational complexity of our model by building a bottleneck type of DIM, as shown in Fig. 3.3(e). On the one hand, we use one single 1×1 convolutional layer in the top to replace the existing individual ones in the different branches for dimensionality reduction. On the other hand, the filter concatenation is replaced by sum-fusion (element-wise addition) which can also help in dimensionality reduction and efficient computation. As a result, this final DIM only needs an additional $27W$ parameters compared to the baseline model. Furthermore, with the help of this computationally-friendly module, our proposed

DINet can reach more than 50 FPS inference time for processing the input images of size 240×320.

In the literature, the atrous spatial pyramid pooling (ASPP) module [2] also utilizes parallel dilated convolutions for learning multi-scale feature representations, as shown in Fig. 3.3(f). In this module, the features extracted at different dilation rates are further processed in separate branches and sum-fused to generate the final results. In contrast, our DIM is just a single plug-in module and its outputs are still features, rather than the final results. Since these two modules share the same idea of using the parallel dilated convolutions, it is also reasonable to use the ASPP module to replace our DIM and its followed decoder network for saliency prediction. Directly insert this ASPP module on the top of DRN cannot guarantee that every pixel in the final results is in the range of [0,1]. We add an extra linear scaling operation after sum-fusion to solve this. ASPP module has two variants: ASPP-S and ASPP-L. The only difference between these two variants is the setting of dilation rates. ASPP-S has smaller dilation rates ($[\alpha, \beta, \gamma, \theta] = [2, 4, 8, 12]$) while ASPP-L has larger rates ($[6, 12, 18, 24]$). The information of these two ASPP-based saliency models is reported in the last two rows in Table 3.1. As observed from this table, with the help of huge extra parameters, the model (DRN + ASPP-S) can obtain similar performance to our DINet. Compared to the ASPP module, our DIM only need one simple decoder network to generate the saliency predictions since we have the sum-fusion before the decoder rather than after it. Another reason for longer inference time in the ASPP-based model is that our DIM performs the 1×1 convolution before the dilated convolutions for dimension reduction while ASPP directly uses dilated convolutions to process these features from DRN. In particular, the difference between the dilated convolutions part of ASPP and our DIM in #parameters is $8 \times 3^2 \times 4 = 288W$ versus $8 + 3^2 \times 3 = 35W$.

Besides, we also investigate other existing multi-scale context feature extraction frameworks, such as image pyramid network (IPN) with shared backbone network and skip-layer network, into our baseline model. The overall comparison among these models are listed in Table 3.1. The extra params (%) term indicates the percentage of the number of additional parameters involved when using this model compared to the baseline model. The best validation loss term means that the smallest loss results of the models evaluated on SALICON validation dataset [3].

TABLE 3.1: Comparison of the baseline model and other models with different multi-scale context feature extraction modules. The model (Baseline+ Inception(e)) in **bold** is our final proposed DINet model.

Model	Total #params	Extra params (%)	Best validation loss	Average Inference time
Baseline (DRN + Decoder)	25.27M	0	0.2776	72.40s
ResNet + Skip-layer + Decoder	26.84M	6.21	0.2793	48.52s
Baseline + Skip-layer			0.2739	76.38s
Baseline + IPN	26.38M	4.39	0.2732	100.54s
Baseline + Inception(a)	30.84M	22.04	0.2720	89.50s
Baseline + Inception(a) - 1×1 Branch	29.72M	17.61	0.2721	87.91s
Baseline + Inception(b)	32.94M	30.35	0.2701	90.22s
Baseline + Inception(c)	29.27M	15.83	0.2696	81.08s
Baseline + Inception(d)			0.2673	81.22s
Baseline + Inception(e)	27.04M	7.00	0.2679	77.39s
DRN + ASPP-S	42.70M	67.98	0.2679	116.92s
DRN + ASPP-L			0.2684	141.38s

The loss function used here is the linear normalization-based total variation distance, as discussed in the next section. The detailed evaluation results corresponding to these loss values are reported in Table 3.4. The average inference time term is the average time of these models for predicting 5,000 validation images with 5 repeats under the same experimental conditions. Among these models in Table 3.1, our DINet achieves a relatively good trade-off between the validation performance and inference speed.

3.2.4 Loss Function

Most saliency models directly predict saliency maps via optimizing loss functions designed for pixel-wise regression/classification. However, the saliency map can be viewed as a probability distribution (PD) of human fixations over the whole image [75]. Pixel-wise prediction, where each pixel is predicted individually, may suffer from the global inconsistency problem as it ignores the spatial relationship between the pixels. Therefore, it is reasonable to use off-the-shelf PD distance metrics as loss functions for training deep saliency models. To convert the predicted saliency maps and their corresponding ground truths into the probability distributions, a normalization method should be applied first. Here, we improve the existing method [75] by replacing their softmax normalization with a simple linear regularization.

Base on the validation experiments in Section 3.3.4, the total variation distance is selected as the final PD loss function for training our DINet. Besides, the unnormalized version of total variation distance is the $\ell 1$-norm loss which is a typical regression loss. Due to these two factors, we use this loss function as an example to illustrate the differences between our proposed linear normalization-based loss function and the existing two types. The total variation distance or $\ell 1$-norm can be broadly formulated by the following equation:

$$L(\mathbf{p}, \mathbf{g}) = \sum_i |p_i - g_i|, \tag{3.2}$$

where $\mathbf{p}$ is the predicted result and $\mathbf{g}$ is the ground truth. The definitions of these two terms are different in each loss function, as listed in the following:

In $\ell 1$-norm (unnormalized loss function),

$$p_i = x_i^p, \quad g_i = x_i^g. \tag{3.3}$$

In softmax normalization-based loss function,

$$p_i = \frac{exp(x_i^p)}{\sum_{i=1}^{N} exp(x_i^p)}, \quad g_i = \frac{exp(x_i^g)}{\sum_{i=1}^{N} exp(x_i^g)}. \tag{3.4}$$

In our linear normalization-based loss function,

$$p_i = \frac{x_i^p}{\sum_{i=1}^{N} x_i^p}, \quad g_i = \frac{x_i^g}{\sum_{i=1}^{N} x_i^g}, \tag{3.5}$$

where $\boldsymbol{x} = (x_1, ..., x_i, ..., x_N)$ is the set of raw saliency response values for either the predicted saliency map ($\boldsymbol{x^p}$) and the ground truth saliency map ($\boldsymbol{x^g}$).

The experimental results in Section 3.3.4 illustrate that our proposed linear normalization-based loss functions perform better than both softmax normalization-based and unnormalized ones. According to the following theorem, for an array whose values range from 0 to 1, the softmax will de-emphasize the maximum values among them while the linear normalization still maintains their initial proportion. Since the ground truth saliency map is an array $\boldsymbol{x^g} \in [0, 1]^N$, the existing loss functions coupled with softmax normalization cannot accurately measure the gaps between the predicted probability distribution and its corresponding ground truth.

Theorem. Given an array $\boldsymbol{x} \in [0, 1]^N$, using Equation (3.4) and Equation (3.5) to normalize this array separately, denote the range of the elements of this two normalized arrays as $[a_s, b_s]$ and $[a_l, b_l]$, respectively. Then, we have:

$$[a_s, b_s] \subset [a_l, b_l].$$

Proof. It is obvious that both these normalization functions are monotonic increasing functions. We also note that $\boldsymbol{x} \in [0, 1]^N$. So, we get the minimum normalized response when $x_i = 0$ and get the maximum when $x_i = 1$. Considering that we have $a_s = \frac{exp(0)}{\sum_i exp(x_i)} = \frac{1}{\sum_i e^{x_i}} > 0 = \frac{0}{\sum_i x_i} = a_l$. Now we only need to prove $b_l \geq b_s$.

In fact, we have:

$$b_l - b_s = \frac{1}{\sum_i x_i} - \frac{e}{\sum_i e^{x_i}} = \frac{\sum_i (e^{x_i} - ex_i)}{\sum_i x_i \sum_i e^{x_i}}.$$

Recall that $x_i \in [0, 1]$, it is easy to prove that $e^{x_i} - ex_i \geq 0$ for every $x_i \in [0, 1]$. So we have $b_l \geq b_s$.

□

3.3 Experiments

In this section, we apply the proposed DINet for visual saliency prediction and report its experimental results on several public saliency prediction benchmark datasets. The effectiveness and efficiency of our method are validated qualitatively and quantitatively.

3.3.1 Saliency Prediction Benchmark Datasets

For evaluating the saliency prediction model, three popular saliency prediction benchmark datasets are adopted. The detailed information of these saliency benchmark datasets are presented as follows:

- **SALICON** [3] contains 10,000 training images, 5,000 validation images, and 5,000 test images, taken from the Microsoft COCO dataset [136]. The spatial resolution of each image in this dataset is 480 × 640. At present, it is the largest public dataset for visual saliency prediction. The ground truths of training and validation datasets are available while the ground truths for the test images are held out. For evaluation on its test dataset, researchers need to submit their results on the SALICON challenge website[3]. Besides, the evaluation protocols and codes are available in the website[4].

- **MIT1003** [4] consits of 1,003 images collected from the Internet. The ground truths for this dataset are created from eye-tracking data of 15 users. The

[3]https://competitions.codalab.org/competitions/3791
[4]https://github.com/NUS-VIP/salicon-evaluation

evaluation codes for this dataset are available in the MIT Saliency Benchmark website[5].

- **MIT300** [10] contains 300 images, including both indoor and outdoor scenarios. The ground truths for this entire dataset are held out. Researchers can only submit the results of their models to the MIT Saliency Benchmark website[5] for evaluation. Currently, the MIT1003 dataset is usually viewd as the training and validation sets for this dataset.

3.3.2 Evaluation Metrics for Saliency Prediction

There are many evaluation metrics to measure the agreement between model predictions and human eye fixations. Following existing works [137, 138], we conduct our quantitative experiments by adopting four widely used saliency evaluation metrics, including AUC (Area Under the ROC curve), sAUC (shuffled AUC), NSS (Normalized Scanpath Saliency), and CC (Linear Correlation Coefficient). For the sake of simplification, we denote the predicted saliency map as P, the ground truth saliency map as G, and the ground truth fixation map as Q. G is obtained by using a Gaussian blur to process Q. The information of these four evaluation metrics is listed in Table 3.2 according to their characteristics.

- **AUC and sAUC**: AUC evaluates the binary classification performance of the predicted saliency map P, where fixation and non-fixation points in its corresponding Q are divided into the positive set and negative set, respectively. By using a specific threshold, P can be classified into the salient and non-salient regions. The ROC curve is obtained by varying this threshold from 0 to 1. Finally, the AUC metric can be calculated by using this ROC curve. Shuffled AUC (sAUC) is introduced to alleviate the influence of center-bias. Differ in AUC, the fixation points of other images in this dataset is used as the negative set in computing sAUC values. However, these two AUC-based metrics have a limitation in penalizing false positives, as reported in [39, 40, 67].

- **NSS** is a specific value-based saliency evaluation metric. This metric is computed by taking the mean of $\bar{P}$ at the human eye fixations Q:

[5]http://saliency.mit.edu/

TABLE 3.2: Summary of saliency evaluation metrics

Metrics	Category	Ground Truth
AUC (area under the ROC curve)	Location-based	Fixation Map (Q)
sAUC (shuffled AUC)	Location-based	Fixation Map (Q)
NSS (Normalized Scanpath Saliency)	Value-based	Fixation Map (Q)
CC (Linear Correlation Coefficient)	Distribution-based	Saliency Map (G)

$$NSS = \frac{1}{N}\sum_{i=1}^{N}\bar{P}(i) \times Q(i), \tag{3.6}$$

where N is the total number of human eye fixations, $\bar{P}$ is the unit normalized saliency map P.

- **CC** is a statistical metric for measuring the linear correlation between two random variables. For saliency prediction evaluation, the predicted saliency maps (P) and ground truth density maps (G) are treated as two random variables. Then, CC is calculated by the following equation:

$$CC = \frac{cov(P, G)}{\sigma(P) \times \sigma(G)}, \tag{3.7}$$

where $cov(\cdot, \cdot)$ and $\sigma(\cdot)$ refer to the covariance and standard deviation, respectively.

3.3.3 Implementation Details

Our model is implemented by using Keras [139]. During training, the weights in Dilated ResNet-50 (DRN) are initialized from the ImageNet-pretrained ResNet-50 model. The weights of the remaining layers are initialized by the default setting of Keras. The whole model is trained with widely used Adam optimizer [140] with an initial learning rate of 10^{-4}. This learning rate will be scaled down by a factor of 0.1 after every two epochs. A mini-batch of 10 images is used in each iteration.

Our DINet is trained with 10,000 training images from the training dataset of SALICON [3]. The validation part of SALICON is used to validate our model. For the MIT1003 dataset [4], we directly use the model trained on the SALICON dataset to evaluate the generalization performance of our model on this dataset.

For testing on the MIT300 dataset [10], we fine-tune our model in the MIT1003 dataset by following the same evaluation protocol in [40, 67]. It is worth mentioning that our DINet can achieve the processing speed as little as 0.02s and 0.03s for one input image of size 240×320 and 320×480, respectively, by using one single GTX 1080 Ti GPU.

3.3.4 Loss Function Analysis

In this subsection, we compare the performance of our baseline models trained by our proposed probability distribution (PD) distance metrics with linear normalization to those trained on standard regression loss functions and softmax normalization based statistical distances. Following [75], some PD distance metrics, including KL divergence (KLD), χ^2 divergence, Total Variation distance (TV distance), Cosine distance, and Bhattacharyya distance, are picked as one of the loss functions in our experiments.

Table 3.3 presents the experimental results for each loss function, as measured by the overall performance with respect to the four aforementioned evaluation metrics on SALICON validation dataset. These results support that: (i) Generally, the loss functions based on PD distance metrics perform better than standard regression loss functions, such as BCE, $\ell 1$-norm, and $\ell 2$-norm in our experiments; (ii) For a specific statistical distance based loss function, our proposed linear normalization method is more compatible than the softmax normalization as it can measure the distance between the predicted PD and its target in a more proper way; (iii) Using NSS loss function alone can obtain an extremely high NSS score while this loss function is not very good at other three evaluation metrics.

The first two observations have been discussed in Section 3.2.4. The reason for (iii) can be illustrated by Table 3.2. NSS is a value-based saliency evaluation metric since it is computed by the average of the normalized saliency values at eye fixation locations. In other words, a saliency map with a higher NSS score is more like a fixation map which is not similar to the fixation density map, i.e. saliency map. Conversely, another three evaluation metrics (CC, AUC, sAUC) prefer the latter one. Therefore, it is difficult to use one single loss function to train the DCNN model for obtaining a promising result on both NSS and other evaluation metrics.

TABLE 3.3: Performance comparison of the baseline models trained with different loss functions on SALICON validation dataset [3].

Loss function	CC	sAUC	AUC	NSS
Linear normalization vs. softmax normalization				
Total Variation distance (linear)	**0.843**	0.788	0.885	3.077
Total Variation distance (softmax)	0.826	0.786	**0.888**	2.906
Bhattacharyya distance (linear)	0.839	0.786	0.881	3.077
Bhattacharyya distance (softmax)	0.828	0.785	0.884	2.992
KLD (linear)	0.842	0.788	0.886	3.070
KLD (softmax)	0.827	0.785	0.884	2.968
χ^2 divergence (linear)	0.826	**0.790**	0.886	2.994
χ^2 divergence (softmax)	0.826	0.786	0.883	2.968
Cosine distance (linear)	0.835	0.789	0.885	3.048
Cosine distance (softmax)	0.828	0.786	0.887	2.999
Standard regression loss				
BCE	0.826	0.785	0.885	2.963
Euclidean (ℓ2-norm)	0.824	0.781	0.884	2.958
ℓ1-norm	0.810	0.783	0.874	2.960
Evaluation metric-based loss				
NSS	0.733	0.782	0.860	**3.411**

3.3.5 Model Visualization

We verify the effectiveness of our DIM by individually visualizing the response of each dilated convolutional branch. This visualization experiment is achieved by adding a new decoder network without non-linear activations at the end of our DIM. Both the additional decoder and the original decoder are jointly trained with the same loss and the same inputs from the DIM. Since this additional decoder is equivalent to a linear operator, the joint decoded output in this decoder can be decoupled into a linear combination of the outputs which are obtained from the individual branches. Moreover, the input dimension of our decoder is the same as the output dimension of every branch in our DIM (all are equal to 256). The individual response of each branch can be easily obtained by feeding this additional decoder with the features learned in this specific branch. By visualizing both joint and individual saliency prediction results, we can analyze the contributions of these dilated convolutional branches in our DIM.

Fig. 3.4 shows the saliency prediction results of five validation images. The first three columns present the saliency maps independently predicted by branch-α, -β

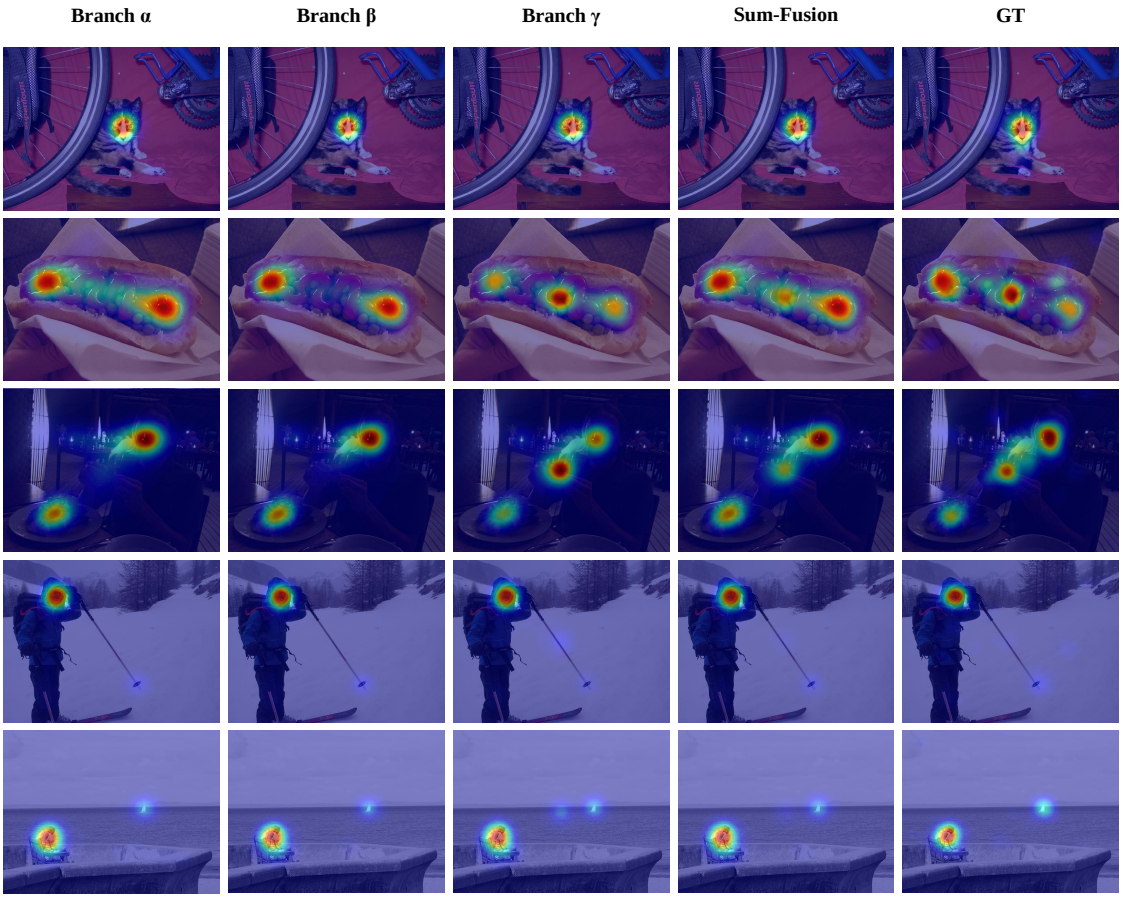


FIGURE 3.4: Influence of each dilated convolutional branch in the DIM to visual saliency. In each column, images are the saliency prediction results by using the features captured from the above indicated branch. GT: Ground Truth.

and -γ, and the fourth column shows the final saliency maps by sum-fusing the outputs of these branches. In general, all of these predicted saliency maps are consistent with the ground truths. Besides, branches with different receptive fields learn to focus on different parts of an input image, as demonstrated in the second and the third rows. In particular, the branch γ, denoted by b_γ, with the largest dilation rate, learns the center-bias implicitly without any additional supervision. These learned center-bias patterns compensate for the negligence on the center salient regions from the other two branches, b_α and b_β, and produce a more accurate saliency prediction result. On the other hand, b_γ sometimes generates false alarms in the center regions with low confidence. In this case, as shown in the last two rows of Fig. 3.4, the previous two branches b_α and b_β can help in reducing this unwanted side-effect on the final fusion results. These three branches in our DIM work in a collaborative manner. The results by using the features from a single branch are no need to be perfect for all possible cases. These incomplete predictions will

TABLE 3.4: Model ablation analysis on SALICON validation dataset [3].

Model	CC	sAUC	AUC	NSS
Influence of backbone network				
ResNet + Decoder	0.776	0.762	0.879	2.456
Baseline (DRN + Decoder)	0.843	0.788	0.885	3.077
Influence of decoder network				
DRN + Decoder(1 Conv layer)	0.838	0.785	0.883	3.052
DRN + Decoder(2 Conv layers)	0.841	0.787	0.884	3.067
DRN + Decoder(4 Conv layers)	0.843	0.788	0.885	3.072
DRN + Decoder(1 Deconv + 1 Conv layers)	0.841	0.787	0.884	3.064
DRN + Decoder(2 Deconv + 1 Conv layers)	0.841	0.788	0.884	3.067
DRN + Decoder(3 Deconv + 1 Conv layers)	0.841	0.787	0.885	3.061
Effectiveness of multi-scale features				
ResNet + Skip-layer + Decoder	0.841	0.786	0.885	3.053
Baseline + IPN	0.849	0.787	0.885	3.086
Baseline + Skip-layer	0.847	0.788	0.886	3.084
Baseline + Inception(a)	0.850	0.788	0.886	3.094
Baseline + Inception(a) - 1×1 Branch	0.849	0.789	0.886	3.091
Baseline + Inception(b)	0.852	0.790	0.886	3.107
Baseline + Inception(c)	0.852	0.790	0.886	3.111
Baseline + Inception(d)	0.854	0.790	0.887	3.114
DINet (Baseline + Inception(e))	0.853	0.789	0.887	3.117
DRN + ASPP-S	0.853	0.789	0.887	3.112
DRN + ASPP-L	0.852	0.789	0.887	3.102
Influence of training image size				
DINet (240 × 320)	0.853	0.789	0.887	3.117
DINet (320 × 480)	0.858	0.790	0.887	3.143
DINet (480 × 640)	0.854	0.789	0.886	3.128
DINet (ensemble)	**0.867**	**0.792**	**0.889**	**3.168**

be ensembled by the sum-fusion to become more comprehensive and reliable final results, which can be also supported by the quantitative results in Table 3.6.

3.3.6 Ablation Study

In this section, we conduct a series of ablation analyses to evaluate the contribution of each component in our DINet on the SALICON validation dataset. The complete ablation results are presented in Table 3.4. It should be noted that all of the models in this table are trained by the proposed linear normalization-based total variation distance loss function.

3.3.6.1 Influence of the backbone network

Our baseline model is built on the DRN where the *output_stride* is equal to 8. As mentioned in Section 2.1.2, the *output_stride* of original ResNet is 32 which means that less spatial information is included in the output of this backbone network and thus leads to unsatisfactory performance. To verify this statement, we compare our baseline model (DRN + decoder) with a more basic model (ResNet + decoder). From the first part of Table 3.4, we can conclude that *output_stride* is one of the key elements for the dense prediction tasks. There is a significant performance gain by replacing the original ResNet with DRN.

3.3.6.2 Influence of the decoder network

In our baseline model, our designed decoder network is just three convolutional layers and one bilinear up-sampling layer in the end. The reason for using three convolutional layers is determined by the experiments. We have tried to use a different number of convolutional or deconvolutional layers before the prediction layer (one convolutional layer followed by a sigmoid activation) to form other decoder networks. These results are reported in the second part of Table 3.4. As we can see that the models with these decoders cannot get good results as our original decoder, i.e. Decoder(3 Conv layers).

3.3.6.3 Effectiveness of multi-scale features

DINet uses the proposed DIM to capture multi-scale contextual features. To support the conclusions in [39, 40, 64] that integrating multi-scale features can further improve saliency detection performance, we incorporate existing alternative multi-scale feature extraction modules, including IPN, skip-layer, inception, and ASPP, into our baseline or backbone network. From the third part of Table 3.4, we can observe that the saliency prediction performance indeed boosted by incorporating the multi-scale features. Especially, when the backbone network is not DRN, the multi-scale features can reduce the performance drop significantly, by comparing two models with the plain ResNet backbone network, i.e. ResNet + Decoder and ResNet + Skip-layer + Decoder. In all these multi-scale saliency prediction frameworks, our proposed inception(d) and (e) obtain the outstanding results among

them. For the reason that inception(e) is more efficient in terms of #parameters and inference time, as illustrated by Table 3.1, we pick this version of DIM to form our DINet.

3.3.6.4 Influence of training image size

The previous experimental results on SALICON validation dataset are all obtained from 240×320 images, whose size is the half resolution of the original SALICON images. Here we want to see the performance of our DINet models which are trained by images with different spatial resolution. From Table 3.4, we find that the DINet trained by input images of size 320×480 can obtain the best performance among these three models. This model will be directly fine-tuned on the MIT1003 dataset for the evaluation of the MIT300 dataset. Note that these evaluation results are the average scores, there are some validation images that perform better in other DINets (240×320 or 480×640). To characterize this phenomenon, we adopt a simple ensemble learning metric, i.e. average voting, to further improve the performance of our model. By using the average results from these three different models, this ensemble model obtains the best scores in our model ablation analysis.

3.3.6.5 Ensemble learning for improving NSS

However, our best model, which is trained by a single total variation distance loss function, still cannot beat two existing works [40, 67] in NSS metrics, as shown in Table 3.5. These two existing models use the NSS itself as one of the loss functions for training. To further improve our performance on NSS metrics, we use the same ensemble learning method as above to combine the results of two DINet models which are trained by using two different loss functions (total variation distance with linear normalization and NSS) separately. The last ensemble model in this table is our final submission to the SALICON test dataset which results in a good comprise between NSS and another three evaluation metrics.

3.3.6.6 Ablation analysis on DIM

We further verify the effectiveness of our DIM by conducting two additional quantitative experiments. In the first experiment, we evaluate the performance of a

TABLE 3.5: Performance comparison of our DINet models trained with different loss functions on SALICON validation dataset [3].

Model	CC	sAUC	AUC	NSS
DINet (TV distance)	**0.867**	**0.792**	**0.889**	3.168
DSCLSTM [40]	0.835	0.788	0.887	3.221
SAM-ResNet [67]	0.844	0.787	0.886	3.260
DINet (NSS)	0.724	0.782	0.861	**3.600**
DINet (ensemble NSS and TV distance))	0.862	**0.792**	0.886	3.310

TABLE 3.6: Dilated inception module ablation analysis within a trained DINet with two decoders on SALICON validation dataset [3].

Type	b_α	b_β	b_γ	CC	sAUC	AUC	NSS
The results on additional decoder network							
0 branch				0.752	0.794	0.858	2.729
1 branch	✓			0.833	0.799	0.882	3.012
		✓		0.811	0.793	0.864	3.025
			✓	0.801	0.759	0.873	2.967
2 branches-sum	✓	✓		0.831	0.799	0.879	3.035
	✓		✓	0.804	0.799	0.869	3.036
		✓	✓	0.813	**0.800**	0.872	3.032
3 branches-sum	✓	✓	✓	**0.853**	0.789	0.886	3.098
The results on original decoder network							
3 branches-sum	✓	✓	✓	**0.853**	0.789	**0.887**	**3.107**

TABLE 3.7: Dilated inception module ablation analysis with individual trained variants of DINet on SALICON validation dataset [3].

Type	b_α	b_β	b_γ	CC	sAUC	AUC	NSS
0 branch				0.843	0.788	0.885	3.077
1 branch	✓			0.847	**0.790**	0.886	3.080
		✓		0.849	0.788	**0.887**	3.086
			✓	0.851	0.788	**0.887**	3.095
2 branches-sum	✓	✓		0.852	0.788	**0.887**	3.099
	✓		✓	0.853	0.788	0.886	3.098
		✓	✓	0.852	0.788	**0.887**	3.103
3 branches-sum	✓	✓	✓	0.853	0.789	**0.887**	**3.117**
3 branches-concat	✓	✓	✓	**0.854**	0.789	**0.887**	3.116

trained DINet with two decoders mentioned in the visualization experiment to investigate the contribution of each dilated convolutional branch in our DIM respectively. In the second experiment, we make a comparison among a set of variants of DINet to explore the impact of the number of parallel dilated convolutional layers.

Table 3.6 shows the results of the first experiment. Each row in this table represents the evaluation results by using the outputs from the indicated branch(es) as the input to a trained decoder. As we can see that, 1 branch type of DIM will learn different bias under its specific receptive fields to help in predicting visual saliency. Specifically, b_α prefers the results with higher sAUC scores, while b_β is more interested in the NSS metric. By comparing the results between the row of 3 branches-sum and the rows in 2 branches-sum type on the first part of this table, we can observe that the performance drop dramatically with the absence of any one branch, which means every branch in our DIM has its irreplaceable impact on the final results. These three branches in our DIM work in a collaborative manner. Even if the performance by using any individual branch is not comparable to the performance of our baseline model, their fused results can deal with the diverse images with different patterns of salient regions. Moreover, the results on the last row show that the features used in the additional decoder can still be decoded by our original decoder with only a little bit of performance drop in the NSS metric. It can guarantee the generality of the above conclusions.

Table 3.7 compares the performance of several variants of DINet. Each row in this table means the evaluation results by testing the individual trained variant which has the indicated branch(es). Especially, the model in the type of 3 branches-sum is the proposed DINet, while the model in 0 branch type is our baseline model. This table shows that using more branches (from 0 to 3), which means using more comprehensive features, will lead to higher performance on evaluation metrics. Besides, in the 1 branch type of DINet, using dilated convolution with a larger dilation rate before the decoder network can achieve a better performance than using a smaller one. It can be credited to the larger size of receptive fields which represent the longer range of dependencies in captured features. Moreover, using concatenation to replace our element-wise addition has a limited impact on the final results, as presented in the last two rows in this table. Mathematically, element-wise addition followed by a convolution layer is a special case of concatenation followed by another convolution layer [141], which can be used to explain this limited difference in evaluation results. In summary, both of these two experiments can verify that the performance gain of our DIM is realized by the corporation of these three parallel dilated convolutional branches.

3.3.6.7 DIM and Global Features

The convolution operation is a typical local operation that computes the responses at a position as a weighted sum of the features at its adjacent positions in the convolutional kernel. As a result, the multi-scale features extracted by our DIM belong to local features. Although our local features have various receptive fields, they cannot replace the global features which can represent the interactions between any two positions, regardless of their positional distance. Recently, non-local block (NLB) [142] is proposed to simulate the non-local operation for the global feature extraction. We can combine this NLB into our DIM for building a more powerful feature refinement module.

Because NLB can maintain the tensor size of the input features, the insertion location of NLB can be flexible. We can use this module as the fourth branch in our DIM. There are two feature fusion ways between this NLB and the three original branches in DIM: sum and concatenation. The results of these two new models and our original DINet are presented in Table 3.8. We can see that using the concat version of the new module can outperform our DIM on CC, sAUC, and NSS three evaluation metrics. These results can be used to validate the effectiveness of global features. The results of the sum version are close to the original results, which means that the non-local features obtained by NLB should be separated from the local features. Simply fusing them cannot obtain a promising performance gain. Besides, we also try to combine the DIM and NLB sequentially. But the results are not as good as the original DINet. It can also verify that the global features obtained by NLB indeed have different properties towards the features learned in DIM.

It should be noted that our proposed DINet [8] still uses the original DIM. This section is newly added to demonstrate that our work can be further improved by incorporating global features.

3.3.7 Performance Comparison

To demonstrate the effectiveness of our proposed DINet model in predicting visual saliency, we quantitatively compare our method with other deep saliency models on SALICON, MIT1003, and MIT300 datasets.

TABLE 3.8: DIM and NLB combination experiment results on SALICON validation dataset [3].

Model	CC	sAUC	AUC	NSS
DINet (DRN + DIM + Decoder)	0.853	0.789	**0.887**	3.117
DRN + sum(DIM,NLB) + Decoder	0.853	0.789	0.886	3.120
DRN + concat(DIM,NLB) + Decoder	**0.855**	**0.790**	**0.887**	**3.127**
DRN + DIM + NLB + Decoder	0.852	0.789	0.886	3.110
DRN + NLB + DIM + Decoder	0.847	0.785	0.885	3.083

TABLE 3.9: Comparison results on SALICON test dataset [3].

Models	CC	sAUC	AUC	NSS
DINet (Ours)	**0.860**	0.782	**0.884**	**3.249**
SAM-ResNet [67]	0.842	0.779	0.883	3.204
DSCLRCN [40]	0.831	0.776	**0.884**	3.157
SAM-VGG [67]	0.825	0.774	0.881	3.143
SalGAN [70]	0.781	0.772	0.781	2.459
SU [92]	0.780	0.760	0.880	2.610
PDP [75]	0.765	0.781	0.882	-
ML-Net [71]	0.743	0.768	0.866	2.789
MxSalNet [72]	0.730	0.771	0.861	2.767
Deep Convnet [143]	0.622	0.724	0.858	1.859
Shallow Convnet [143]	0.562	0.658	0.821	1.663
DeepGazeII [69]	0.479	**0.787**	0.867	1.271

Table 3.9 shows the evaluation results on SALICON test dataset. The results of other models come from their papers or the leaderboard of this dataset. In this table, the results in bold indicate the best performance method on each evaluation metric. As it can be observed, our DINet outperforms all competitors on CC, AUC, and NSS three metrics. The DeepGazeII [69] model gets the best sAUC score and relatively lower scores on other metrics. However, the saliency maps generated by this model actually are very blurred/hazy and visually different from the ground truths, as shown in the left part of Fig. 3.5. This is because AUC-based metrics mainly relied on true positives without significantly penalizing false positives [39, 67].

The results on MIT1003 are reported in Table 3.10. We directly use the DINet trained on SALICON dataset to evaluate the generalization performance of our model on the whole MIT1003 dataset, as done in [38]. Our model also achieves

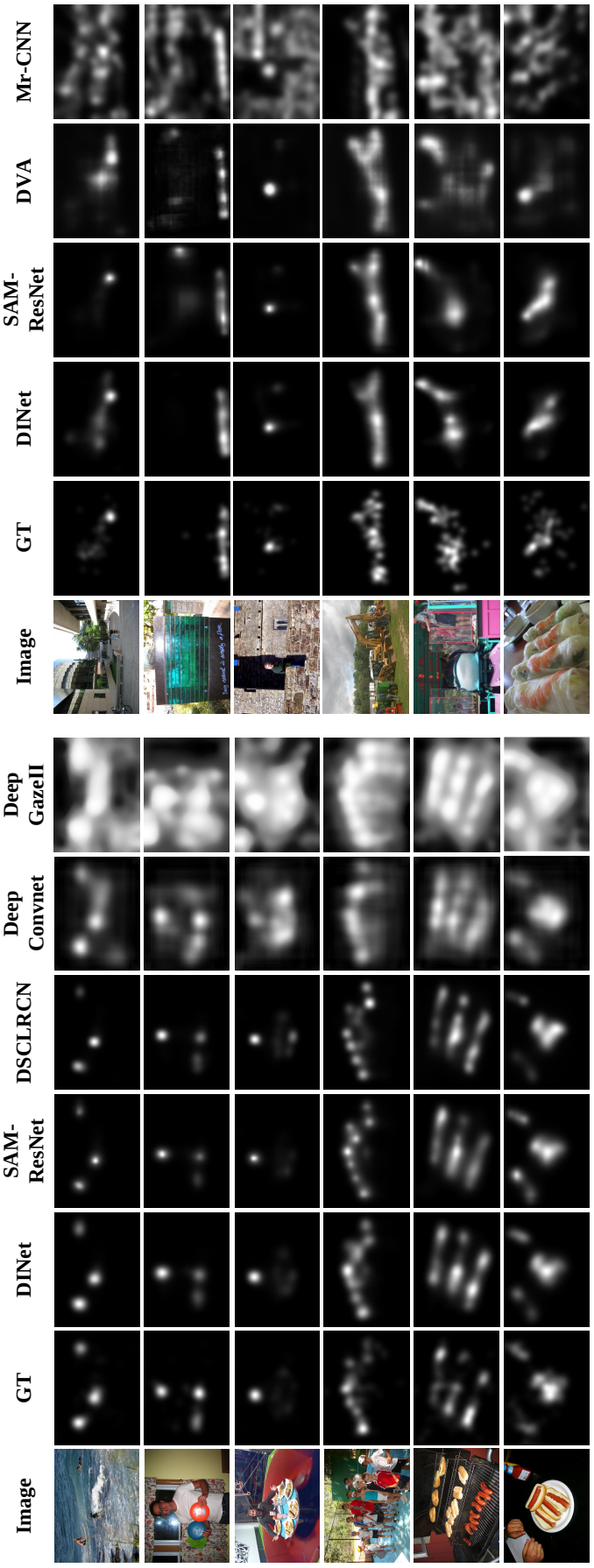


FIGURE 3.5: Qualitative comparison results on two datasets. Left images are from SALICON validation dataset [3], while right images are from MIT1003 dataset [4]. GT: Ground Truth.

TABLE 3.10: Comparison results on MIT1003 dataset [4].

Model	CC	sAUC	AUC	NSS
DINet (w/o finetune)	**0.67**	0.70	**0.88**	**2.40**
DVA [38]	0.64	**0.77**	0.87	2.38
GBVS [50]	0.42	0.66	0.83	1.38
eDN [62]	0.41	0.66	0.85	1.29
Mr-CNN [144]	0.38	0.73	0.80	1.36
BMS [51]	0.36	0.69	0.79	1.25
ITTI [49]	0.33	0.66	0.77	1.10

TABLE 3.11: Comparison results on MIT1003 validation dataset [4].

Model	CC	sAUC	AUC	NSS
DINet (w finetune)	**0.87**	**0.77**	**0.91**	**3.27**
SAM-ResNet [67]	0.77	0.62	**0.91**	2.89
SAM-VGG [67]	0.76	0.61	**0.91**	2.85
DeepFix [39]	0.72	0.74	0.90	2.58
DINet (w/o finetune)	0.67	0.72	0.89	2.50

TABLE 3.12: Comparison results on MIT300 dataset [10].

Model	CC	sAUC	AUC	NSS
DSCLRCN [40]	**0.80**	0.72	0.87	**2.35**
DINet (Ours)	0.79	0.71	0.86	2.33
SAM-ResNet [67]	0.78	0.70	0.87	2.34
DeepFix [39]	0.78	0.71	0.87	2.26
SAM-VGG [67]	0.77	0.71	0.87	2.30
SALICON [64]	0.74	**0.74**	0.87	2.12
SalGAN [70]	0.73	0.72	0.86	2.04
PDP [75]	0.70	0.73	0.85	2.05
DVA [38]	0.68	0.71	0.85	1.98
ML-Net [71]	0.67	0.70	0.85	2.05
SalNet [143]	0.58	0.69	0.83	1.51
BMS [51]	0.55	0.65	0.83	1.41
DeepGazeII [69]	0.52	0.72	**0.88**	1.29
GBVS [50]	0.48	0.63	0.81	1.24
Mr-CNN [144]	0.48	0.69	0.79	1.37
eDN [62]	0.45	0.62	0.82	1.14
ITTI [49]	0.37	0.63	0.75	0.97

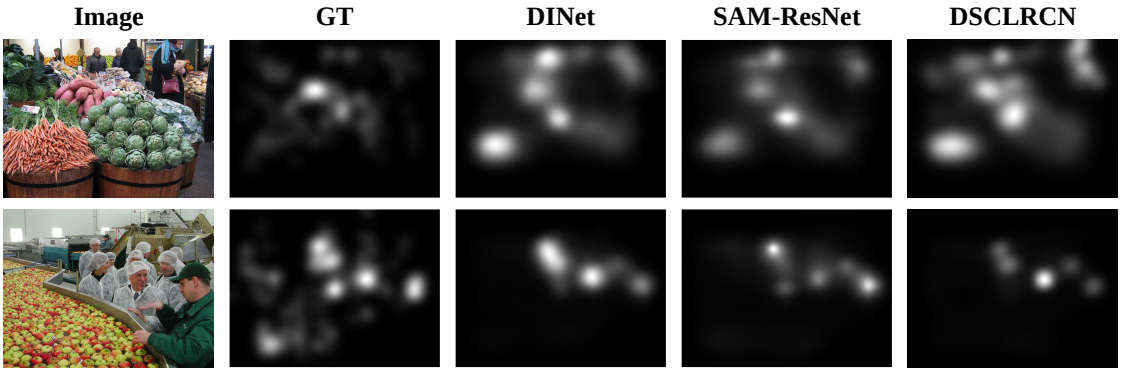


FIGURE 3.6: Some failure cases of our DINet and two competitors. Images are from SALICON validation dataset [3].

promising results on this dataset which verifies its robustness and generality. Qualitative comparison results of our model with other deep saliency models on SALICON validation and MIT1003 datasets can be found in Fig.3.5. This figure can also support that our results match the ground truth saliency maps best among all the compared models in both two datasets.

For the evaluation on MIT300 dataset, we first fine-tune our DINet in the MIT1003 dataset. The fine-tuned results are shown in Table 3.11, As we can see, the performance of our model improves significantly after fine-tuning which can also outperform other existing fine-tuned models. The results on the MIT300 dataset are presented in Table 3.12. Different in the previous two datasets, our DINet can not outperform the DSCLRCN model [40]. Our model may over-fitted on MIT1003 dataset which leads to lower generalization performance on the MIT300 dataset. Both the DSCLRCN model and our DINet use multi-scale features to further improve saliency prediction performance. Besides, the DSCLRCN model incorporates the global context and scene context by using a spatial LSTM [145] method and additional Places-CNN [146] backbone network to achieve this performance. Consequently, their model is more complex and much slower than our method. When predicting visual saliency on an image with a size of 480×640, the DSCLRCN model needs 0.27s while our DINet needs only 0.06s.

However, despite the good results, there are still a small number of failure cases, as shown in Fig. 3.6. These bad cases are caused by the fact that so many objects are cumulated in a single image. Within them, the relative importance of these objects cannot be fully learned by simply utilizing the multi-scale contextual features without a higher level of visual understanding. Therefore, some non-salient regions are

highlighted (like the first row) or some salient regions are missed, as shown in the second row. Note that SAM and DSCLRCN models suffer from the same problem as ours. It can be concluded that even the state-of-the-art deep saliency models still cannot fully understand the relative importance of image regions in such semantically rich scenes. To further approach human-level performance, saliency models will need to discover increasingly higher-level concepts in images for determining an appropriate amount of visual attention on a certain image region.

3.4 Summary

In this work, we have proposed a dilated inception network for visual saliency prediction. The multi-scale saliency-influential factors are captured by an efficient and effective dilated inception module. In particular, a new visualization method specially designed for DIM is proposed to verify its effectiveness. The whole model works in a fully convolutional encoder-decoder architecture, which is trained end-to-end and lightweight for time-efficiency. Furthermore, we adopted a set of linear normalization-based probability distribution distance metrics as loss functions to formulate the saliency prediction task as a probability distribution prediction problem. With such loss functions, our models can perform better than those trained by using either standard regression loss functions or existing softmax normalization-based probability distribution distance metrics. Experimental results on the challenging saliency prediction benchmark datasets have demonstrated the outstanding performance of our model concerning other relevant saliency methods. In the next chapter, we will present our second work for salient object detection.

Chapter 4

Progressive Self-Guided Loss for Salient Object Detection

4.1 Introduction

Salient object detection (SOD)[1] aims to segment the entire salient foreground objects from the background [31]. It is an important pre-processing step for many object-level computer vision applications, such as object detection and recognition [22, 25], image editing and manipulating [26, 27], visual tracking [28], semantic segmentation [29] and image retrieval [147].

Early SOD methods [81, 83, 90] mainly rely on hand-crafted features and heuristic clues to separate foreground and background regions. However, due to the lack of high-level semantic information, these methods are unreliable when detecting salient objects in cluttered and complex scenes. Lately, convolutional neural networks (CNNs), especially the fully convolutional networks (FCNs) [68], lead the recent advances in SOD [85]. Owing to the powerful capacity of extracting high-level semantic information, these FCN-based SOD methods have shown superior performance than conventional methods. However, their predicted saliency maps still suffer from incomplete predictions, as shown in Fig. 4.1. We can observe that even the best existing works still cannot uniformly detect the entire salient objects. Their predictions contain several miss-detected or untrustworthy detected regions, like 'holes', within the salient objects. The key issue is that strong appearance

[1] The work in this chapter is submitted as [107].

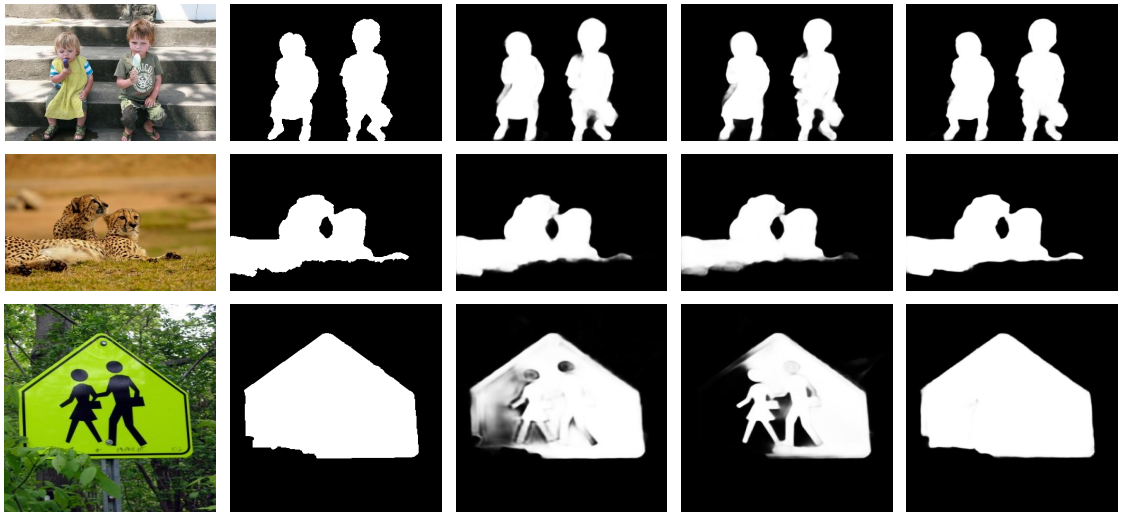


FIGURE 4.1: Visual examples of our method and two relevant existing methods (best viewed digitally with zooming). GT means the ground truth saliency map annotated by humans. Results generated by PoolNet [5] and EGNet [6] suffer from the problem of incomplete predictions. More examples are presented in Fig. 4.5.

changes may happen in the interiors of the salient objects. A common way to address this problem is to find more discriminate feature representations and effective feature aggregation strategies [5, 6, 106].

In this work, we present a new way to address the above-mentioned problem by investigating the training loss. Most FCN-based SOD models use the binary cross-entropy (BCE) as their training loss. But BCE loss is a typical pixel-wise loss function which only accounts for the pixel-wise difference between labels and predictions, ignoring the spatial dependencies of salient object pixels. Models trained with BCE loss usually have the problem of incomplete predictions since every pixel is predicted individually. Therefore, a more suitable training loss is required. Several efforts [41–43] have been made along this direction. However, their proposed losses are not specifically designed for capturing the spatial dependencies among salient pixels. In this work, we focus on progressively modifying the training supervision to create a Progressive Self-Guided (PSG) loss. Unlike the existing works which utilize the labeled saliency maps only or consider additional labels from other related tasks [5, 43, 85], we propose to further process the current network predictions for creating a series of new auxiliary training supervisions in the loss function. The principal idea is that the training process of a SOD model can be decomposed

into several steps. For each step, this model will be provided with some feasible training targets for reducing the training difficulty. As such, its outputs can be progressively optimized during this step-wise training. Specifically, a morphological closing operation, which can help to remove small holes inside the foreground objects, is applied to the current network predictions to generate the new auxiliary training supervisions as part of the overall loss function. The obtained auxiliary training supervisions are similar but more complete than the current-stage network predictions, hereby providing some incentives to the SOD model for approaching them. More importantly, these newly created training targets are keeping refined from the progressively optimized network predictions during training. In such a manner, the spatial dependencies of salient object pixels are implicitly characterized. Consequently, the SOD model can be guided by these progressive supervisions to highlight more complete salient objects step-by-step, even trained with the simple BCE loss.

Besides the progressive supervisions, we also propose a new multi-scale feature aggregation module (MS-FAM) to capture and aggregate the multi-scale features adaptively. In this module, the local context information at different scales is extracted by using parallel dilated convolutions [8, 148] and then sum-fused by applying a branch-wise attention mechanism to characterize their respective importance adaptively. To demonstrate its effectiveness in SOD, we build an encoder-decoder network equipped with these MS-FAMs. In particular, the encoder network is adapted from the feature pyramid network (FPN) [109] architecture where multiple MS-FAMs are inserted to achieve the adaptive multi-scale feature aggregation for further improvement.

Our DINet [8], presented in Chapter 3, is also an FCN-based architecture. It means it can be directly applied to SOD. However, according to our validation experiments in Section 4.3.6, DINet is not good as existing works [5, 6] which are specifically designed for SOD. The main reason is that our DINet is a lightweight architecture, which is suitable for a simpler dense prediction task like eye fixation prediction. While for SOD, a more complex yet effective architecture is needed to defeat other counterparts. Therefore, we propose a new SOD architecture in this work instead of keeping using DINet.

The performance of our proposed SOD model is evaluated on six widely used benchmark datasets. The peer comparison results indicate that our model can achieve

state-of-the-art performance with the help of our proposed PSG loss. Meanwhile, the PSG loss can be directly applied to train other existing SOD models without architecture modification for better alleviating their incomplete prediction problem.

In this work, our contributions can be summarized as follows:

- We propose a novel progressive self-guided (PSG) loss to alleviate the problem of incomplete predictions in the existing SOD models. To the best of our knowledge, this self-guided loss is the first attempt to supervise the SOD model with its own intermediate predictions. As such, the progressive and auxiliary training supervisions are created for step-wisely guiding the training process.
- We propose to apply a simulated morphological closing operation on the network predictions to generate the above auxiliary training supervisions. As a result, the spatial dependencies of salient object pixels are progressively characterized since these generated supervisions are progressively expanded from their sources. Consequently, such progressive training supervisions can guide the SOD models to highlight more complete salient objects step-by-step.
- A new multi-scale feature aggregation module is proposed to build our SOD architecture for further improvement. Benefiting from this module, our SOD architecture takes full advantage of adaptive multi-scale feature aggregation to locate and detect salient objects effectively.

4.2 Progressive Self-Guided Loss and Our SOD Architecture

In this section, we first present the main idea of our progressive self-guided (PSG) loss. Fig. 4.2 gives a simplified illustration of our training losses. Then, an overview of our proposed SOD architecture is described.

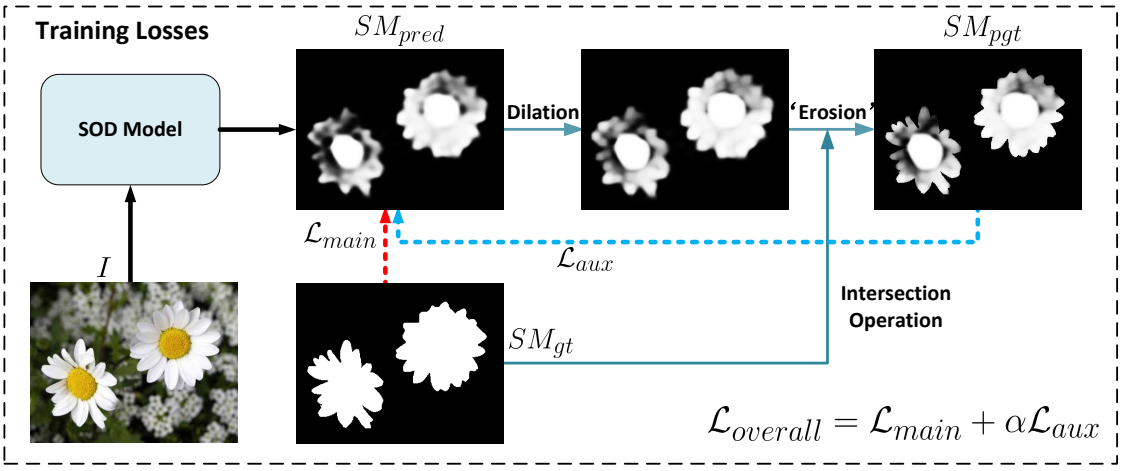


FIGURE 4.2: An illustration of our training losses. In PSG loss ($\mathcal{L}_{aux}$), the predicted saliency map (SM_{pred}) is firstly morphological dialted to expand the boundaries of the detected regions and fill the 'holes' within them, and then morphological eroded by using the intersection operation with the ground truth (SM_{gt}) to obtain a correct and more complete progressive training supervision (SM_{pgt}).

4.2.1 Progressive Self-Guided Loss

4.2.1.1 Motivation and Formulation

Current supervised learning frameworks for SOD use the pairs of the input images and their corresponding labeled saliency maps for training. The training loss of these SOD models mainly focuses on computing the difference between the network predictions and the labeled saliency maps. In other words, given a set of N training images I and the corresponding ground truth saliency maps SM_{gt}, the training loss $\mathcal{L}_{main}$ used by existing works can be described by:

$$\mathcal{L}_{main} = L(SM_{pred}, SM_{gt}) = L(\mathcal{M}(I;\theta), SM_{gt}), \tag{4.1}$$

where $SM_{pred} = \mathcal{M}(I;\theta)$ represents the predicted saliency maps SM_{pred} obtained by feeding the input images I to a SOD model $\mathcal{M}$ under the parameter setting θ. $L(\cdot,\cdot)$ indicates one of the loss computation formulas.

However, as discussed in Section 2.2.3, no suitable loss function $L(\cdot,\cdot)$ can exactly describe the spatial dependencies of salient object pixels in the SM_{gt}, which results in the problem of incomplete predictions. Compared with the efforts of designing more suitable loss, the investigations on the training targets are seldom

investigated. Some of the recent works [6, 99, 106] applied deep supervision by utilizing the SM_{gt} to guide the intermediate predictions. But the performance gain of this technique is not obvious as there is lacking guidance towards characterizing the spatial dependencies. Our idea is to decompose the training process of a SOD model into several steps. For each step, this SOD model will be provided with feasible and step-wise training targets for exploring the spatial dependencies. Such progressive and auxiliary training targets (SM_{pgt}) can be generated by further processing the current network predictions (SM_{pred}). The desirable auxiliary training targets should be similar but more complete than the network predictions for providing some incentives for approaching them. In a nutshell, our PSG loss can be described by:

$$\mathcal{L}_{aux} = L(SM_{pred}, SM_{pgt}) = L(SM_{pred}, f(SM_{pred})), \tag{4.2}$$

where $f(\cdot)$ denotes a kind of processing method used for generating the SM_{pgt}. Note that, the same $L(\cdot, \cdot)$ is used in this auxiliary loss function as the L_{main} for simplification.

The proposed PSG loss is an auxiliary loss that cannot be used as the sole loss for training the SOD models. If the SM_{pred} are filled by zeros, it will be hard to make the SM_{pgt} different from them. In this case, PSG loss will be trapped in the zero value, consequently leading to zero gradients. Therefore, PSG loss should be coupled with a normal training loss $\mathcal{L}_{main}$ for training the SOD models. Therefore, the overall loss is formulated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{main} + \alpha\mathcal{L}_{aux}, \tag{4.3}$$

where α is a non-negative parameter which is used to control the relative importance of the PSG loss. The remaining parts of this section cover the choice of the loss computation formula $L(\cdot, \cdot)$ and the implementation of the processing method $f(\cdot)$.

4.2.1.2 Hybrid Loss Computation

Following [43], we apply a hybrid loss to compute the difference between the network predictions and one of the training targets. This hybrid loss can be defined

as:

$$L(\cdot,\cdot) = L_{bce}(\cdot,\cdot) + L_{dice}(\cdot,\cdot), \tag{4.4}$$

where $L_{bce}(\cdot,\cdot)$ and $L_{dice}(\cdot,\cdot)$ denote the BCE and Dice loss [42], respectively. The detailed computation formula for these two loss are as follows:

$$L_{bce}(X,Y) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot log(x_i) + (1-y_i)\cdot log(1-x_i))], \tag{4.5}$$

$$L_{dice}(X,Y) = 1 - \frac{2\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i}, \tag{4.6}$$

where X represents the one of the predicted results SM_{pred}, Y is the corresponding SM_{gt} or SM_{pgt}, and N is the total number of pixels in X or Y.

BCE loss can help with the convergence of all pixels, regardless of their labels. Dice loss is used to measure the overlap degree between X and Y. By taking this loss into consideration, our SOD model can obtain a better F-measure score. Our validation experiments show that using this hybrid loss instead of using any individual loss in these two choices can achieve better performance.

4.2.1.3 Morphological Closing Operation and Our Simulated Version

Morphology refers to a set of image processing operations that process images based on shapes [149]. In a morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors by applying a structuring element as the comparison window. There are two basic morphological operations: dilation and erosion. The former adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The closing operation means dilation followed by erosion operation which can be used in removing small holes inside the foreground objects. This closing operation can be directly applied to the results of existing SOD models presented in Fig. 4.1 as a kind of post-processing methods for completing them. However, it is time-consuming to perform this post-processing for every prediction as the optimized size of the structuring element is not unified in the test datasets.

We prefer to embed this post-processing into the training loss for creating the progressive training supervisions to teach the SOD models.

The morphological dilation operation can be directly replaced by the max-pooling operation where the kernel in the max-pooling is exactly the same as the structuring element in the dilation operation. However, the erosion operation does not have ready-made alternatives. The erosion operation is used to shrink the object regions which are enlarged by the dilation operation. If we discard the erosion operation, some of the regions in the SM_{pgt} in the later training epochs will lie outside of the SM_{gt}, which results in wrong training guidance. After careful consideration, we decide to use the intersect operation between the network predictions and the SM_{gt} as the approximate alternative function of the erosion operation. There are two advantages to adopt this approximated operation: on the one hand, using SM_{gt} to intersect with the network predictions can always maintain the relationship of $SM_{pgt} \subseteq SM_{gt}$, which is essential to keep the correct supervision. Using the strict closing operation to process the SM_{pred} may break the above relationship and lead to some side effects. On the other hand, the information of the SM_{gt} is transparent to the SOD models during the training stage. Benefiting from this information, the predicted regions in the early training epochs are enlarged without any shrinkage in a reasonable margin for advancing the training convergence. To sum up, our simulated morphological closing operation can be described by:

$$f(SM_{pred}) = e(d(SM_{pred})) \approx maxpool(SM_{pred}) \cap SM_{gt}, \tag{4.7}$$

where dilation operation $d(\cdot)$ is equal to the max-pooling operation $maxpool(\cdot)$ and erosion operation $e(\cdot)$ is approximated by the intersect operation with the SM_{gt}.

Fig. 4.3 presents a visual example to illustrate the effectiveness of our PSG loss. For the incomplete regions in SM_{pred}, the dilation operation in PSG loss expand them in a region growing manner for guiding the training in the next epoch. While for the incorrectly predicted pixels, the intersection operation can help to avoid wrong guidance. In such PSG loss, the SOD model will be encouraged to detect the mis-detected pixels in the neighboring regions of the current predictions with a higher priority. This is because the loss penalty weight in these pixels is $(1+\alpha)$ rather than 1 in other mis-detected yet separate pixels. As a result, the spatial dependencies of salient object pixels are characterized in our PSG loss and the

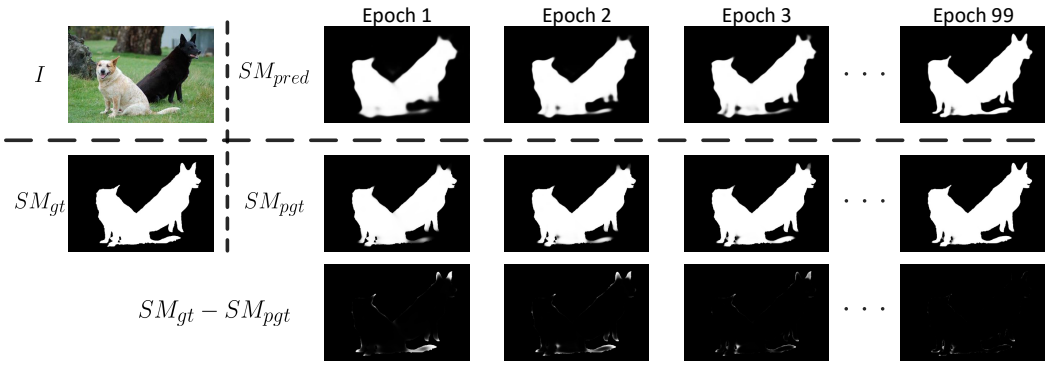


FIGURE 4.3: A visual example to show the epoch-wise difference between the SM_{pred} and SM_{pgt} in the PSG loss. The results of the models from the first three epochs and the last epoch are presented.

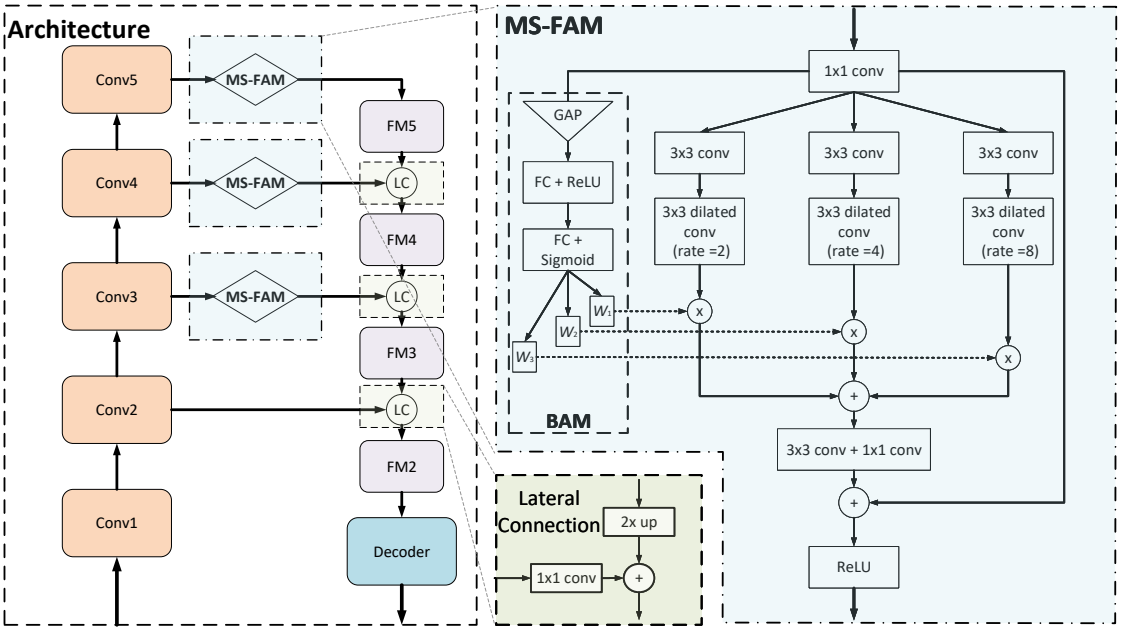


FIGURE 4.4: An overall framework of our proposed SOD model. GAP and FC are the abbreviation of global average pooling and fully-connected layer, respectively. FM5 denotes the group of feature maps with the same spatial size as the output of Conv5, and so on.

SOD model can be guided to fit the training samples in such a progressive region-growing way. Moreover, the progressive manner can be partially reflected by the similarity of the current SM_{pgt} and SM_{pred} in the next epoch and the shrinkage of the residual maps between the current SM_{pgt} and SM_{gt} in Fig. 4.3.

4.2.2 Architecture Overview

An overview of our proposed SOD architecture is depicted in Fig. 4.4. Our model consists of three key components: feature pyramid network (FPN) [16], multi-scale feature aggregation module (MS-FAM), and decoder network.

4.2.2.1 Feature Pyramid Network

We apply an FPN-like framework as the encoder network for feature extraction. The construction of the FPN involves a bottom-up pathway, a top-down pathway, and lateral connections. The bottom-up pathway (from Conv1 to Conv5) is the stage for multi-level feature extraction. The output of the last layer of each convolutional block (except Conv1) in this pathway will be used as the source of feature maps for the feature aggregation in the top-down pathway (from FM5 to FM2) via using lateral connections (LCs). Each LC aggregates feature maps of the same level (spatial size) from the bottom-up pathway and the top-down pathway. Instead of making multi-level predictions like in the original FPN, we only use the final aggregated feature maps in FM2 to feed the decoder network for saving the inference time.

4.2.2.2 Multi-scale Feature Aggregation Module

To explicitly incorporate the multi-scale features for further improvement, we insert several multi-scale feature aggregation modules (MS-FAMs) before the LCs in our encoder network. In MS-FAM, the local context information at different scales is extracted by using parallel dilated convolutions with different dilation rates. For enhancing the discrimination ability of the aggregated features, we propose to apply a branch-wise attention mechanism (BAM) to characterize the respective importance of these multi-scale local features adaptively. It is achieved by adapting a SE block [150] from modeling the channel-wise feature dependencies to weighting the importance of each dilated branch individually. Specifically, the outputs of this attention mechanism are three separate scalars (W_1, W_2, W_3) in the range of 0 to 1. These branch-wise scalars are the learned weights for their corresponding dilated branches to make the sum-aggregated features more discriminative and adaptive.

Moreover, a shortcut connection [66] is applied at the end of our MS-FAM to make it easier to optimize.

In general, the function of our MS-FAM is to further refine the feature maps obtained from the convolutional blocks in FPN for generating more powerful multi-scale feature maps of the same tensor size. Since our MS-FAM can maintain the tensor size of the input features, the insertion location of this module can be flexible in a fully convolutional network (FCN). In this work, three individual MS-FAMs are inserted at three connection paths between Conv3-5 and FM3-5, as depicted in Fig. 4.4.

This MS-FAM can be viewed as an advanced version of our dilated inception module (DIM) presented in Chapter 3. The major difference between these two modules is the usage of BAM. It means that the effectiveness of MS-FAM can be proved by our DIM and MS-FAM can be used in visual saliency prediction with great potential.

4.2.2.3 Decoder Network

The decoder network in our framework is used to convert the aggregated features into saliency maps. Our basic FCN-based decoder network is built by stacking six convolutional layers and one bilinear up-sampling layer in the end. In addition, we can incorporate our MS-FAMs into this basic decoder network for building a more powerful one. The detailed configuration for this part is summarized in Sec. 4.3.3.

4.3 Experiments

4.3.1 SOD Datasets

To evaluate the performance of our method, we conduct extensive experiments on six widely used SOD benchmark datasets: ECSSD [37], PASCAL-S [151], DUT-OMRON [90], HKU-IS [86], SOD [152], and DUTS [101]. The detailed information of these six datasets is presented as follows:

- **ECSSD** [37] contains 1,000 images with semantically meaningful but structurally complex natural contents. The pixel-wise ground truths are annotated by five subjects.
- **PASCAL-S** [151] is composed by 850 images selected from the validation dataset of PASCAL VOC 2010 [94].
- **DUT-OMRON** [90] consists of 5,168 images of relatively complex backgrounds and high content variety. Each image is carefully labeled by five users.
- **HKU-IS** [86] contains 4,447 images with complex scenes that typically contain multiple salient objects with relatively diverse spatial locations.
- **SOD** [152] consists of 300 challenging images from the Berkeley segmentation dataset [153]. Most of the images in this dataset have multiple salient objects with low contrast.
- **DUTS** [101] is currently the largest SOD benchmark. The images in this dataset are divided into two non-overlapping subsets: DUTS-TR and DUTS-TE. The former one contains 10,553 images designed for training and the latter has 5,019 test images. The training images are selected from the ImageNet DET train/val dataset [36], and the test images from the ImageNet test dataset [36] and the SUN dataset [154]. Following the recent works [5, 6, 103, 104, 106], we use the DUTS-TR dataset for training our SOD models and keep the remaining datasets for test.

4.3.2 Evaluation Metrics for SOD

There are three widely adopted evaluation metrics, including precision-recall (PR) curves, F-measure and mean absolute error (MAE), for evaluating the performance of SOD models.

- **Precision-Recall (PR)** is calculated based on the comparison between the binarized saliency map and the ground truth:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{4.8}$$

where TP, FP, FN denote true-positive, false-positive, and false-negative, respectively. To get the binary saliency map, a set of thresholds ranging from 0 to 255 is applied to the raw saliency map, each threshold will produce a pair of Precision-Recall value to form a PR curve for visualizing the model performance.

- **F-measure**, denoted as F_β, comprehensively considers both precision and recall by computing the weighted harmonic mean of them:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{4.9}$$

where β^2 is empirically set to 0.3 to weight more on precision. Instead of reporting the whole F-measure plot, the maximum F_β (maxF) values are reported as done in recent works.

- **MAE** is used to measure the average pixel-wise absolute error between the predicted saliency map $SM_{pred} \in [0,1]^{W \times H}$ and its ground truth $SM_{gt} \in \{0,1\}^{W \times H}$:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |SM_{pred}(i,j) - SM_{gt}(i,j)|. \tag{4.10}$$

4.3.3 Implementation Details

Our SOD model and PSG loss are implemented in Pytorch. The backbone network used in this paper is ResNet-50. The input images to our models are all resized to 352×352 for both training and test. The α used in computing the total loss is set to 1 for equally treating $\mathcal{L}_{main}$ and $\mathcal{L}_{aux}$. The feature dimension in our MS-FAMs is fixed to 64. Our decoder network consists of three stacked MS-FAMs, two 3×3 convolutional layers, one 1×1 convolutional layer for the final prediction, and one bilinear up-sampling layer in the end. In the basic one, these MS-FAMs are degraded by 1×1 convolutional layers with ReLU.

During training, we apply the random horizontal flipping to the training images for data augmentation. The weights in the backbone network are initialized from its ImageNet pre-trained model. The weights of the remaining layers are initialized by the default setting of Pytorch. Our models used in the experiments are trained with

TABLE 4.1: Performance comparison on six widely used SOD datasets. The symbols ↑ and ↓ denote that a score being larger and smaller is better, respectively. In each column, the best three results are marked in red, green, and blue, respectively.

Model	Backbone	ECSSD [37]		PASCAL [151]		DUT-O [90]		HKU-IS [86]		SOD [152]		DUTS-TE [101]	
		maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓
VGG-based													
DCL [91]	VGG-16	0.900	0.078	0.853	0.113	0.804	0.086	0.907	0.055	0.818	0.193	0.823	0.148
UCF [96]	VGG-16	0.912	0.070	0.863	0.115	0.810	0.120	0.907	0.061	0.836	0.164	0.838	0.111
Amulet [97]	VGG-16	0.916	0.060	0.871	0.100	0.825	0.097	0.913	0.050	0.819	0.141	0.846	0.084
NLDF [98]	VGG-16	0.907	0.065	0.857	0.098	0.798	0.079	0.908	0.048	0.823	0.123	0.842	0.065
DSS [99]	VGG-16	0.892	0.055	0.816	0.093	0.776	0.063	0.894	0.040	0.751	0.121	0.840	0.056
PAGRN [105]	VGG-19	0.907	0.064	0.851	0.089	0.767	0.071	0.907	0.047	0.767	0.145	0.850	0.055
ResNet-based													
SRM [100]	ResNet-50	0.919	0.056	0.870	0.084	0.812	0.069	0.913	0.046	0.821	0.126	0.858	0.058
PAGENet [85]	ResNet-50	0.920	0.046	0.870	0.076	0.833	0.062	0.917	0.036	0.797	0.110	0.869	0.051
PiCANet [104]	ResNet-50	0.933	0.048	0.889	0.075	0.840	0.064	0.925	0.044	0.840	0.103	0.886	0.050
DGRL [103]	ResNet-50	0.919	0.043	0.871	0.075	0.808	0.062	0.911	0.036	0.761	0.103	0.858	0.049
BASNet [43]	ResNet-34	0.922	0.040	0.870	0.076	0.841	0.057	0.918	0.033	0.778	0.112	0.876	0.047
CPD [106]	ResNet-50	0.933	0.040	0.881	0.071	0.826	0.056	0.922	0.033	0.820	0.110	0.878	0.043
PoolNet [5]	ResNet-50	0.932	0.042	0.884	0.075	0.837	0.055	0.929	0.032	0.844	0.100	0.892	0.040
EGNet [6]	ResNet-50	0.937	0.041	0.886	0.074	0.840	0.053	0.929	0.031	0.846	0.097	0.896	0.039
Ours	ResNet-50	0.940	0.035	0.896	0.062	0.844	0.054	0.934	0.028	0.828	0.097	0.896	0.037

Adam optimizer with an initial learning rate of 5×10^{-5}. This learning rate will be scaled down by a factor of 0.1 after half of the training epochs. The batch size is set to 20 and the total number of training epoch is 99. It is worthy to mention that the average inference time of our method is 0.015s to process an image of size 352×352 by using a GTX 1080Ti GPU. The source codes of our method will be made publicly available.

4.3.4 Performance Comparison

We compare our method with 14 recent FCN-based SOD models: DCL [91], UCF [96], Amulet [97], PAGRN [105], DSS [99], NLDF [98], PiCANet [104], SRM [100], PAGENet [85], DGRL [103], BASNet [43], CPD [106], PoolNet [5] and EGNet [6]. Among them, the former six models are VGG-based while the latter eight models are ResNet-based. For fair comparisons, all the saliency maps of the above methods are released by the authors or generated by using the available source codes. The evaluation codes are adopted from [5, 6, 99] with the necessary rectification (We note that their original codes will output the precision and recall in the wrong order and thereby return a wrong F_β result).

In Table 4.1, we present the quantitative performance comparison results on the six SOD datasets. For maxF and MAE results on each dataset, the best three models are highlighted in red, green, and blue, respectively. From this table, we

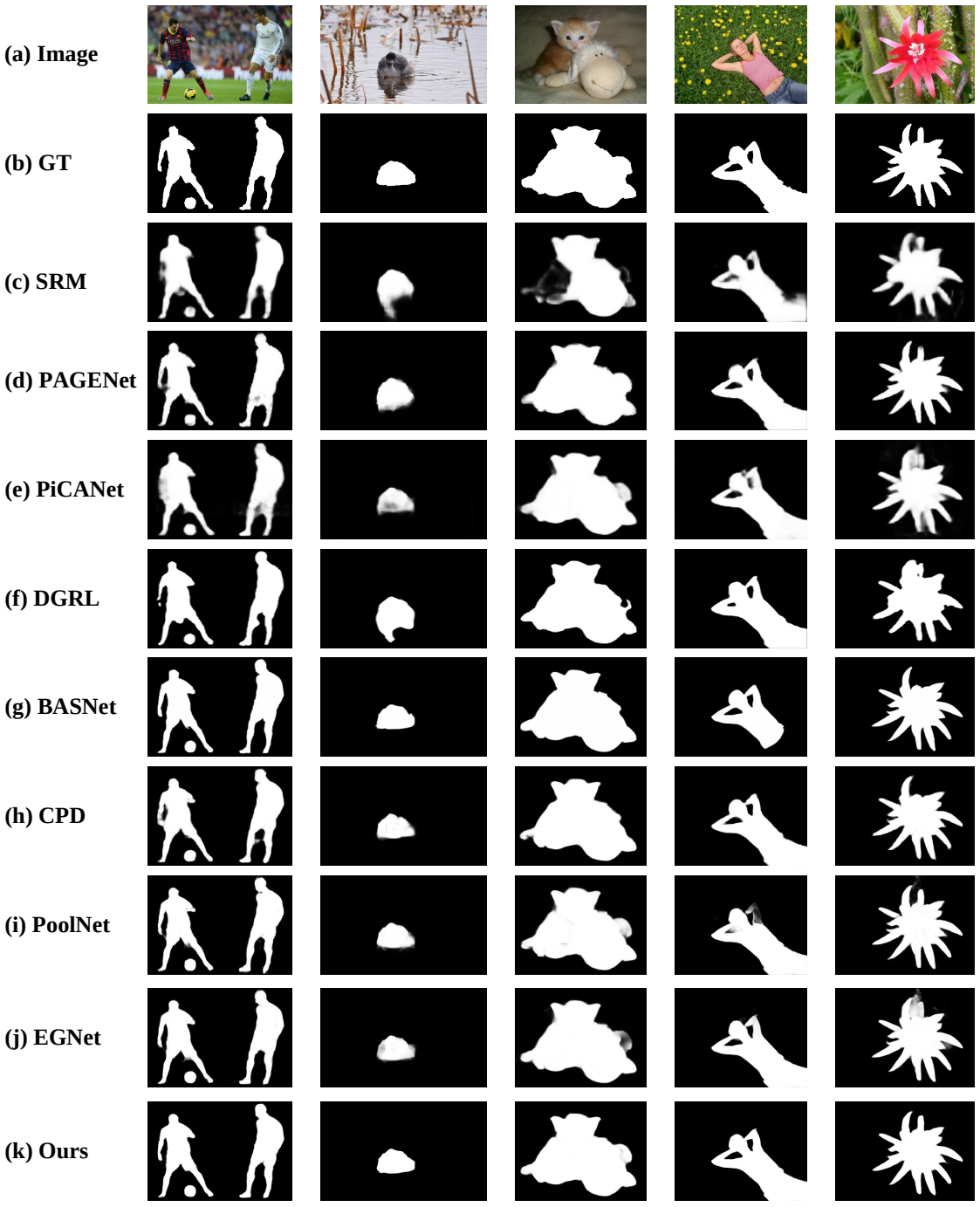


FIGURE 4.5: Visual comparisons of our method and other ResNet-based models on some representative examples (best viewed digitally with zooming).

can observe that our method achieves state-of-the-art performance on almost all benchmark datasets, except on the SOD dataset [152]. Fig. 4.5 shows a visual comparison of the results of our method against other ResNet-based models. As we can see that the detected salient objects by using our method are more complete than other competitors.

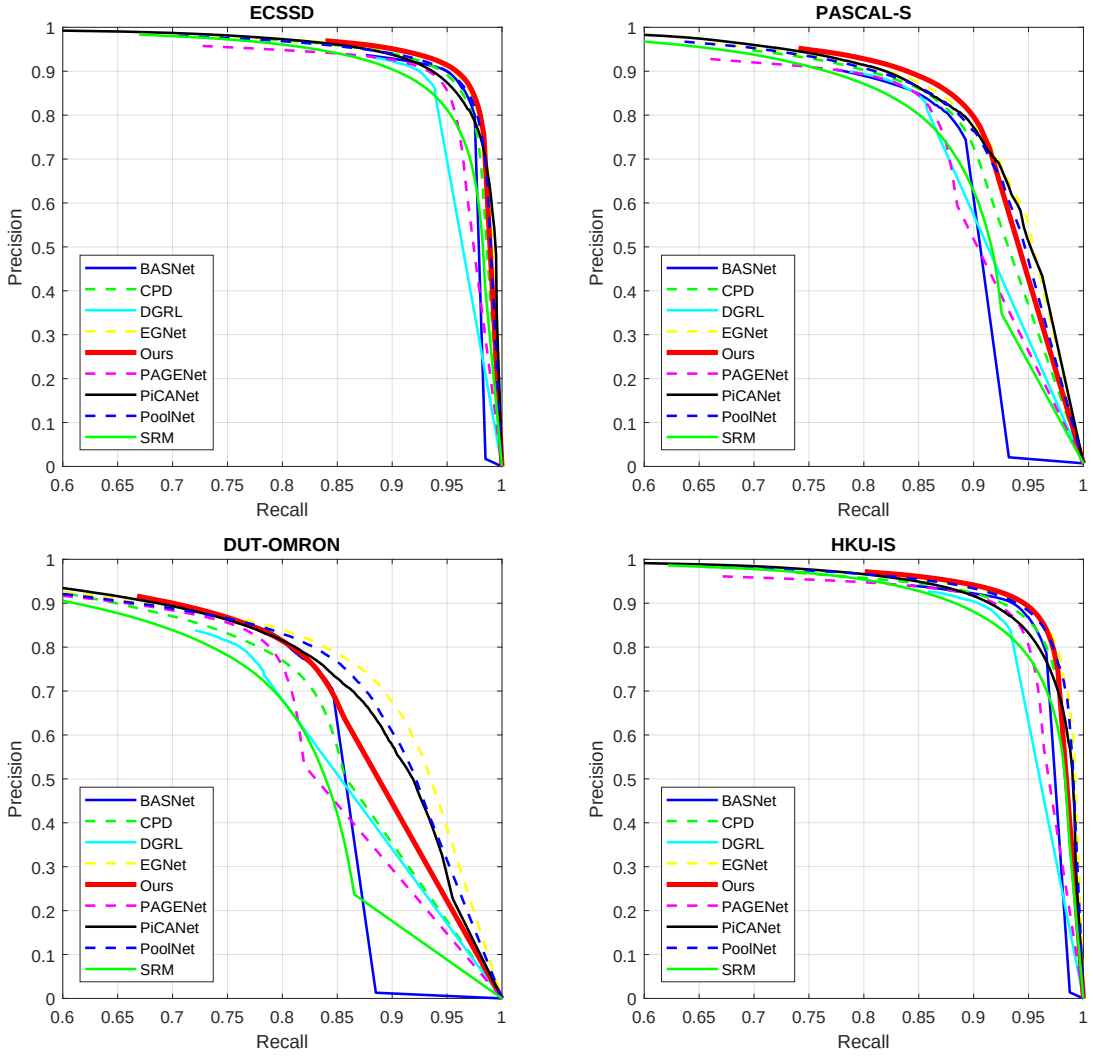


FIGURE 4.6: Performance comparison with PR-curve on four SOD benchmarks. Ours model obtains promising performance on three of them. Best viewed in color.

Besides, the precision-recall curves of ResNet-based methods on the first four datasets are provided in Fig. 4.6. We can see that our method outperforms its counterparts on ECSSD, PASCAL-S, and HKU-IS datasets. While for the DUT-O dataset, although our model has the best maxF value, its PR curve is not promising on this dataset. It should be emphasized that this drawback mainly results from our architecture. If we plot the PR curve of our model without trained with PSG loss on the DUT-O dataset, we will obtain a worse curve.

TABLE 4.2: Model ablation analysis of our method with maxF (higher is better) and MAE (lower is better) on six SOD benchmarks. In each column, the best two results are marked in red and green, respectively.

Model Variants		PSG Loss	ECSSD [37]		PASCAL [151]		DUT-O [90]		HKU-IS [86]		SOD [152]		DUTS-TE [101]	
MS-FAMs in Encoder	MS-FAMs in Decoder		maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓
Effectiveness of MS-FAM														
			0.926	0.043	0.883	0.067	0.816	0.058	0.925	0.031	0.807	0.111	0.880	0.040
	✓		0.930	0.042	0.881	0.068	0.818	0.055	0.922	0.031	0.820	0.109	0.881	0.039
✓			0.930	0.038	0.890	0.066	0.831	0.054	0.930	0.029	0.830	0.101	0.887	0.038
✓	✓		0.936	0.038	0.888	0.065	0.830	0.054	0.930	0.028	0.819	0.104	0.890	0.036
Effectiveness of PSG Loss														
		✓	0.926	0.042	0.877	0.068	0.817	0.055	0.926	0.030	0.812	0.111	0.881	0.039
	✓	✓	0.930	0.040	0.884	0.064	0.824	0.053	0.926	0.030	0.805	0.107	0.884	0.038
✓		✓	0.935	0.036	0.890	0.063	0.832	0.054	0.930	0.029	0.807	0.102	0.891	0.036
✓	✓	✓	0.940	0.035	0.896	0.062	0.844	0.054	0.934	0.028	0.828	0.097	0.896	0.037

4.3.5 Ablation Study

In this subsection, we conduct a series of ablation experiments to analyze the contribution of two key components, including the multi-scale feature aggregation module (MS-FAM) and progressive self-guided (PSG) loss, in our method. The quantitative results of our ablation study on the six SOD benchmarks are summarized in Table 4.2. Our baseline model, as shown in the first row of this table, consists of an FPN-like encoder and a fully convolutional decoder network without MS-FAMs. Our MS-FAM is a flexible convolutional module for extracting and fusing the multi-scale features, which can be applied in both the encoder and decoder network. As a result, we have four model variants as the subjects in this ablation study.

4.3.5.1 Effectiveness of MS-FAM

As shown in Table 4.2, the models equipped with MS-FAMs can easily outperform the baseline model in most of the test datasets, regardless of the application location of MS-FAMs. These results verify the conclusion in many SOD papers [86, 97, 99] that the SOD performance indeed boosted by incorporating the multi-scale features.

Moreover, the model with MS-FAMs used in the encoder network can achieve better results in all test benchmarks than the one with MS-FAMs used in the decoder. It means that multi-scale features should be better incorporated into the encoder network before the saliency inference stage. The model in the fourth row of Table 4.2 is the final architecture of our SOD model. It can be used to

illustrate that making a powerful decoder network can also help in improving the SOD performance.

The branch-wise attention mechanism (BAM) plays an important role in our MS-FAM. If we discard it by only applying the parallel dilated convolutions with sum-aggregation in MS-FAM, the performance of our final architecture will drop by 1.8% in the SOD dataset and 1.0% in the PASCAL-S dataset, in terms of maxF. As a result, our model will be surpassed by some existing works in Table 4.1.

4.3.5.2 Effectiveness of PSG Loss

PSG loss can be applied to the above-mentioned model variants for guiding their training. From Table 4.2, we can observe that the model trained with PSG loss can surpass its normally trained version in most benchmarks, convincingly demonstrating the effectiveness of this loss. We note that there are several 'abnormal' values in the SOD dataset [152] that some of the model variants trained with PSG loss will obtain a worse result. We guess that the SOD dataset only has 300 images for testing which may have a different label distribution from the training images in the DUTS-TR dataset. Comparing Table 4.1 with 4.2, our SOD model cannot achieve the state-of-the-art performance without the help of the PSG loss.

The kernel size of max-pooling used in our PSG loss is one of the most important hyper-parameters for designing it. The current setting is using a 3×3 max-pooling layer for the morphological dilation. We find that the performance of using 5×5 or even larger size of max-pooling is not good as using this one, but still better than not using PSG loss. Obviously, using a larger size of max-pooling will result in larger shape changes between the SM_{pred} and SM_{pgt}. One possible explanation is that the difference by using a larger size of max-pooling is out of the learning capacity of our SOD model in one training epoch which makes it hard to optimize.

4.3.5.3 PSG Loss with other training losses

Besides the hybrid loss (BCE + Dice) used in the existing experiments, our PSG can work with other training losses as well. We choose the fourth model variant (Model4), i.e. the model with MS-FAMs in both encoder and decoder, in our

ablation study as the experiment subject in this part. Five commonly used segmentation training losses, such as $\ell 1$-norm, $\ell 2$-norm, KLD, Dice, and BCE loss, are individually used to validate the generalization ability of PSG loss. The results are presented in Table 4.3. For each loss coupled with our PSG loss, the improved or degraded results are marked in red or green, respectively. We can easily find that the performance of Model4 trained with a specific training loss can be boosted by incorporating the PSG loss.

TABLE 4.3: Performance comparison of the training losses and the losses combined with our PSG loss. For each loss coupled with our PSG loss, the improved or degraded results are marked in red or green, respectively.

Loss	PSG Loss	ECSSD [37]		PASCAL [151]		DUT-O [90]		HKU-IS [86]		SOD [152]		DUTS-TE [101]	
		maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓
$\ell 1$-norm		0.920	0.038	0.861	0.065	0.797	0.054	0.920	0.029	0.766	0.100	0.867	0.038
	✓	0.925	0.038	0.875	0.062	0.811	0.052	0.921	0.029	0.762	0.105	0.878	0.036
$\ell 2$-norm		0.927	0.042	0.878	0.067	0.815	0.056	0.923	0.033	0.821	0.105	0.877	0.041
	✓	0.933	0.042	0.881	0.069	0.824	0.056	0.927	0.032	0.836	0.103	0.885	0.040
KLD		0.935	0.041	0.889	0.067	0.829	0.054	0.929	0.032	0.837	0.103	0.891	0.038
	✓	0.937	0.041	0.892	0.067	0.835	0.056	0.930	0.032	0.838	0.105	0.892	0.040
Dice		0.919	0.039	0.872	0.065	0.797	0.053	0.920	0.029	0.730	0.101	0.871	0.038
	✓	0.924	0.037	0.886	0.069	0.840	0.060	0.922	0.028	0.754	0.098	0.883	0.041
BCE		0.935	0.040	0.886	0.066	0.828	0.054	0.931	0.031	0.841	0.102	0.888	0.037
	✓	0.937	0.039	0.891	0.065	0.836	0.054	0.933	0.030	0.833	0.104	0.889	0.039
BCE + Dice		0.936	0.038	0.888	0.065	0.830	0.054	0.930	0.028	0.819	0.104	0.890	0.036
	✓	0.940	0.035	0.896	0.062	0.844	0.054	0.933	0.028	0.828	0.097	0.896	0.037

4.3.5.4 Post-processing vs. PSG Loss

We also try to use the normal morphological closing operation as the post-processing method for refining the results of our models. Specifically, the fourth model variant, i.e. the model with MS-FAMs in both encoder and decoder, in our ablation study is chosen as the subject. We find that the evaluation performance of using a unified small kernel size of the closing operation, such as 3×3, 5×5, and 7×7, on all of the testing images, are roughly the same as the results of non-using. But it doesn't mean that the morphological closing post-processing is useless. In fact, there is an optimal kernel size for refining a specific image. This kernel size should be dependent on the size of its incomplete regions. As shown in the first row images in Fig. 4.7, using a small kernel of closing cannot complete the big 'holes' predicted by our subject model. For this type of image, a larger kernel is more suitable. For other images in Fig. 4.7, simply using a larger kernel of closing will worse the performance by wrongly merging the non-salient pixels into the salient regions. It usually occurs in the images contain some salient objects with clear but close boundaries. Moreover, closing operations cannot rectify false positives. In

TABLE 4.4: Performance comparison of the original models and the models retrained with our PSG loss. The symbol + denotes the retrained one. For each retrained model, the improved and degraded results are marked in red and green, respectively.

Model	ECSSD [37]		PASCAL [151]		DUT-O [90]		HKU-IS [86]		SOD [152]		DUTS-TE [101]	
	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓	maxF↑	MAE↓
CPD	0.933	0.040	0.881	0.071	0.826	0.056	0.922	0.033	0.820	0.110	0.878	0.043
CPD+	0.933	0.039	0.884	0.067	0.834	0.055	0.925	0.032	0.826	0.106	0.883	0.042
EGNet	0.937	0.041	0.886	0.074	0.840	0.053	0.929	0.031	0.846	0.097	0.896	0.039
EGNet+	0.937	0.041	0.888	0.074	0.842	0.054	0.931	0.031	0.852	0.101	0.900	0.038
Ours	0.934	0.038	0.888	0.065	0.830	0.054	0.930	0.028	0.819	0.104	0.890	0.036
Ours+	0.940	0.035	0.896	0.062	0.844	0.054	0.934	0.028	0.828	0.097	0.896	0.037
DINet	0.929	0.044	0.881	0.069	0.825	0.058	0.923	0.034	0.830	0.100	0.881	0.040
DINet+	0.932	0.041	0.885	0.067	0.830	0.058	0.926	0.032	0.822	0.101	0.883	0.042

this case, using a larger kernel of closing is also a bad choice. By comparing the results of using post-processing and those of using PSG loss, we can see that the accuracy of raw prediction is more important than the choice of the post-processing parameter.

To summarize, the performance of using the closing operation as post-processing is not as good as conducting it in the loss function. There are two main drawbacks for the former manner: firstly, the post-processing result is sensitive to the choice of kernel size. In PSG loss, a fixed and small kernel can be adopted as the model is progressively guided and the testing result can be iteratively optimized through the epochs. Secondly, the post-processing does not always yield better results. In PSG loss, correct auxiliary supervisions can be always guaranteed with the prior knowledge of SM_{gt}.

4.3.6 Application in Existing Methods

Our proposed PSG loss is an auxiliary loss function that can be directly applied in training any end-to-end SOD models without architecture modification. In this paper, we try to apply our PSG loss into two recent SOD methods: CPD [106] and EGNet [6], for evaluating the generalization ability of this loss. These two models are retrained by using their released source codes as well as the default training settings. In Table 4.4, we report the quantitative results of the original models and the models retrained with our PSG loss. We can see that PSG loss can further improve the performance of the CPD and EGNet models on almost all SOD benchmarks.

Since SOD is still a dense prediction task, our proposed DINet [8], which is presented in Chapter 3, can be retrained for detecting salient objects as well. From Table 4.4, we can observe that our PSG loss can also help our DINet in improving its performance on SOD. Although DINet only requires 0.007s to process an image of size 352×352, the performance gap between DINet and this new work is not marginal. SOD is more complex than the visual saliency prediction task as the labeled saliency maps in SOD is more dense and structured. Even though DINet has a more powerful backbone network, i.e. Dilated ResNet-50 (DRN), its feature aggregation module and decoder network are at a distinct disadvantage against our new model. Owing to its simple architecture, DINet cannot obtain state-of-the-art performance in SOD. This is the main reason why we propose a new model for SOD instead of keeping using DINet.

4.4 Summary

In this work, we have proposed a simple yet effective progressive self-guided (PSG) loss for assisting the training of deep learning-based salient object detection (SOD) models. Our PSG loss simulates the morphological closing operation on the model predictions for creating progressive and auxiliary training supervisions epoch-wisely. The effectiveness and the generalization ability of this loss have been validated in our experiments. Moreover, we also propose a new multi-scale feature aggregation module (MS-FAM) to build our SOD models. Experimental results on six widely used SOD benchmark datasets have demonstrated the outstanding performance of our method with respect to other FCN-based models. In the future, we will consider promoting our PSG loss into other dense prediction tasks and investigate other post-processing techniques into this loss for further improvements. In the next chapter, we will present our last work for saliency-guided image quality evaluation.

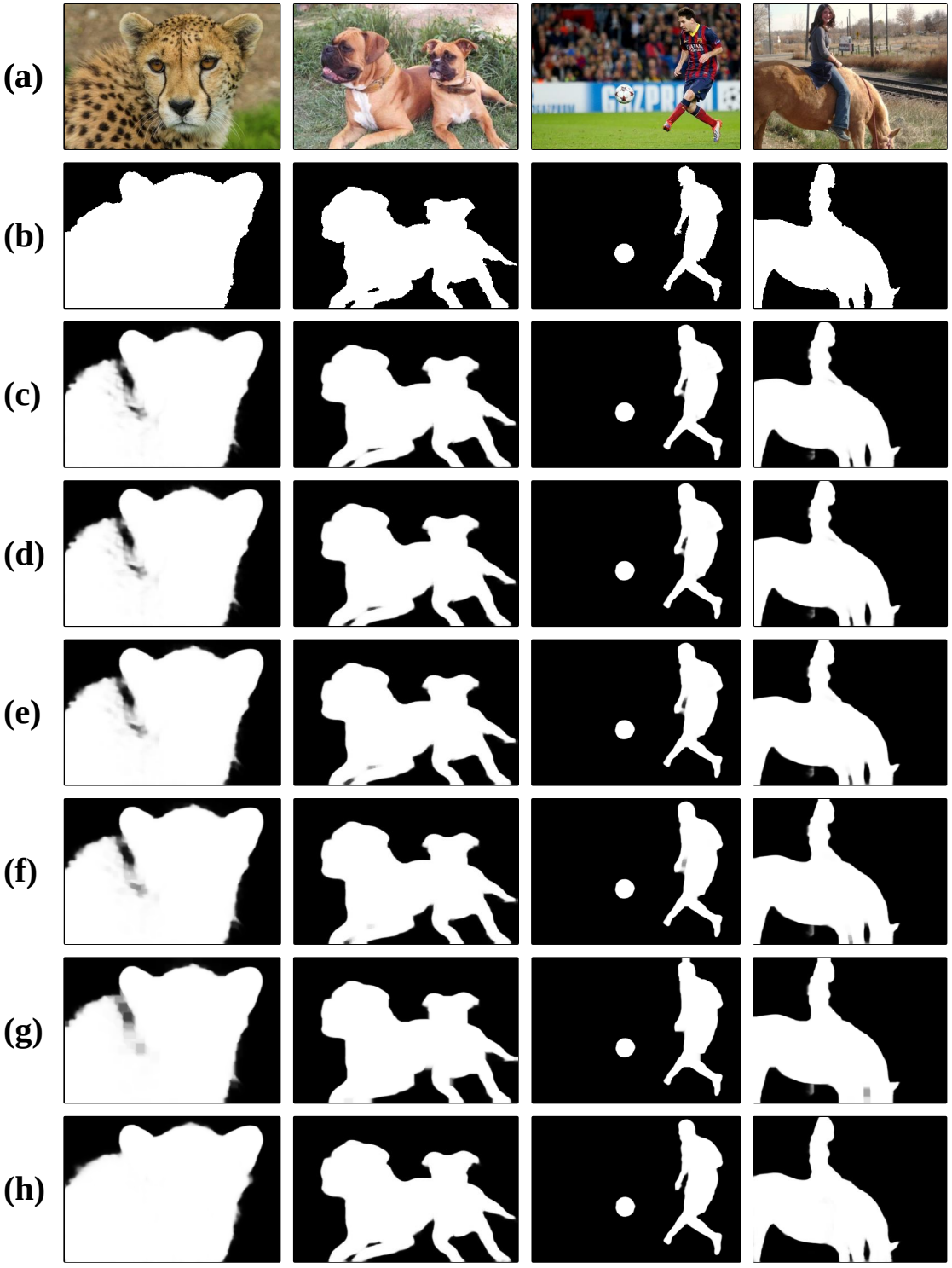


FIGURE 4.7: Visual comparisons of our model with different kernel sizes of closing operation and PSG Loss (best viewed digitally with zoom). (a) Image, (b) GT, (c) Model4, (d)-(g) Model4 + 3×3, 5×5, 7×7, and 13×13 closing, respectively. (h) Ours (Model4 + PSG Loss). Model4 denotes the fourth model variant in Table 4.2.

Chapter 5

Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment

5.1 Introduction

Image quality assessment (IQA)[1] aims to evaluate the perceptual quality of a digital image in a manner that is consistent with human subjective opinions. It is a fundamental problem in many perceptual-based visual media applications, such as image deblurring [156, 157], image compression [158], image super-resolution [159], and more. According to the accessibility of the pristine reference images, IQA models can be classified into full-reference (FR) [110–113], reduced-reference (RR) [114–116], and no-reference (NR) [18, 117–120] three types. Among them, NR-IQA has a broad range of application scenarios since reference images are not accessible in most practical applications, especially for evaluating the qualities of real-world images with authentic distortions.

Traditional NR-IQA methods generally follow a two-stage processing pipeline including feature extraction and quality regression. Related works have shown that the performance of these NR-IQA models heavily depends on their carefully designed quality-aware features based on the domain knowledge of natural scene

[1]The work in this chapter has been published in [155].

statistics (NSS) [121–123] and human visual properties [117, 160]. Lately, with the advent of deep convolutional neural network (CNN), these hand-crafted feature-based NR-IQA models are surpassed by deep CNN-based models due to the powerful capacity of deep CNN architectures in jointly optimizing the feature extraction and quality regression modules in a data-driven manner.

To learn a better feature representation, some recently proposed deep IQA models seek to use a multi-task learning strategy where an auxiliary sub-task and a primary sub-task (i.e., quality evaluation) are jointly optimized in an end-to-end manner. With the help of such an auxiliary yet closely related sub-task, these models can learn more discriminant feature representations from the input raw data to improve their quality evaluation performance effectively. For example, the recent MEON model in [45] considered distortion identification as the auxiliary sub-task. This sub-task intends to make use of the distortion category information which is available in the legacy IQA datasets with some common synthetic distortions [11, 161, 162]. However, the massive Internet images captured by real cameras are usually afflicted by complex mixtures of multiple authentic distortions [7, 163], which cannot be well-simulated by the limited algorithm-generated distortions in these legacy IQA datasets. As a result, such a distortion identification sub-task cannot accurately identify the complex mixtures of distortions existing in authentically distorted images and may lead to performance degradation when applying to evaluate the real-world images with diverse authentic distortions.

To address the above-mentioned limitation of the existing multi-task deep CNN-based NR-IQA models, we propose to use visual saliency prediction to replace their distortion identification as the auxiliary sub-task for providing more universal yet closely related perceptual information to facilitate quality evaluation. Compared with the distortion identification sub-task, visual saliency always exists when viewing every image, regardless of its distortion type. More importantly, related work [113] has reported that saliency information is highly correlated with image quality. The rationale is that human beings tend to focus on visually salient areas while assessing image quality [17, 126, 127]. This inspires us to incorporate the visual saliency prediction as the auxiliary sub-task to learn a powerful multi-task deep IQA model for the quality evaluation on the authentically distorted images. Fig. 5.1 presents some Internet images along with the estimated saliency maps by using our proposed saliency prediction model–DINet [8]. Images (a)-(c) are obviously

FIGURE 5.1: Examples of Internet images and their saliency maps with different image quality levels. The images in the first row are from KonIQ-10k dataset [7], and larger MOS (mean opinion score) shown in the bottom indicates better subjective perceptual quality. Their saliency maps in the second row are generated by our DINet [8] and fused with the original images where a pixel with brighter intensity indicates a higher probability of attracting human visual attention.

with low visual quality due to the severe distortions/blurs on the salient regions. Image (d) is slightly better yet still has distortions on its attended areas. The quality of image (e) is not as good as image (f) because the second most salient object (i.e., the man in the rear) is not that clear.

The proposed saliency-guided deep neural network (SGDNet) inherits the same idea of other saliency-based NR-IQA methods [127, 133] by estimating saliency maps as a kind of local weighting functions to measure the local visual importance dependence for facilitating quality evaluation, but it differs greatly from these two relevant works in two main aspects. On the one hand, their immediate visual saliency targets are optimized only with the single global quality scores as supervisions while these targets in our SGDNet have the direct supervisions, provided by our proposed saliency model [8] in the Chapter 3. On the other hand, the outputs of our saliency prediction sub-task are transparent to the quality evaluation task by working as a kind of spatial attention priors/masks on the extracted features from the whole image for feature fusion. As a result, our method is an end-to-end image-based approach that avoids using the problematic local patch quality scores as labels in the training process.

The performance of our SGNet is evaluated on several publicly available IQA benchmark datasets. The peer comparison results indicate that our SGDNet can achieve state-of-the-art performance on both authentically and synthetically distorted IQA datasets. Meanwhile, the ablation study shows that the quality evaluation performance is indeed boosted by incorporating saliency information and our multi-task

learning framework can further improve the performance due to its learned adaptive spatial attention priors for better perceptually-consistent feature fusion.

In this work, our contributions can be summarized as follows:

- We propose an end-to-end optimized SGDNet to incorporate learnable saliency information into the challenging NR-IQA task. To the best of our knowledge, it is the first attempt to optimize the saliency prediction and quality evaluation sub-tasks together in an end-to-end multi-task learning framework for alleviating the overfitting problem. The proposed SGDNet is particularly suitable to blindly evaluate the perceptual qualities of real-world images with authentic distortions.
- Our SGDNet is trained with more informative labels including saliency maps and global quality scores simultaneously for better quality evaluation. More importantly, the learned saliency information from the saliency prediction sub-task is transparent to the primary quality regression sub-task by providing a kind of adaptive spatial attention priors for the perceptually-consistent feature fusion.

5.2 Saliency-Guided Deep Neural Network

5.2.1 Overview

We propose an end-to-end multi-task saliency-guided deep CNN model for NR-IQA. It consists of two sub-tasks including visual saliency prediction and image quality evaluation which are jointly optimized with a shared feature extractor. In the literature, there are some IQA datasets [129, 164, 165] that provide saliency maps, in addition to the MOS (mean opinion scores), to investigate the interaction of visual saliency and quality evaluation. However, the scale of these datasets is too small for training a powerful saliency prediction model. Considering that, our saliency prediction sub-network is trained with proxy saliency maps produced by a teacher saliency model–our DINet [8], which are trained on the large-scale saliency prediction dataset–SALICON [3]. To verify the effectiveness of these proxy saliency maps, we implement another saliency-guided deep CNN model, called

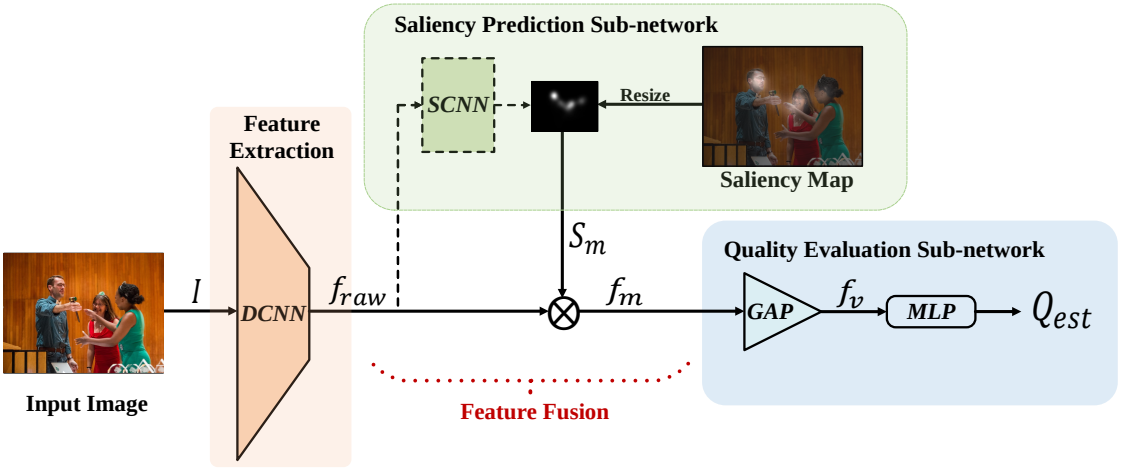


FIGURE 5.2: Architectures of two variants of proposed saliency-guided deep CNN models. (1) Direct SGDNet (without the saliency prediction sub-network indicated by dashed lines); (2) (Multi-task) SGDNet: use a saliency prediction sub-network to predict saliency map under the supervision of target saliency map and then incorporate this learned saliency map with the extracted features to evaluate the image quality. Definitions of notations used in this figure are described in Sections 5.2.2 and 5.2.3.

Direct SGDNet, which directly uses the proxy saliency maps as the additional model input to provide a kind of saliency guidance for improving the accuracy of quality evaluation. The architectures of these two models are depicted in Fig. 5.2. For these two models, either the proxy or learned saliency maps are served as the spatial attention priors to fuse the extracted features from the whole input images. As such, the visual importance dependence of local regions over the whole image is modeled and a perceptually-consistent feature fusion is achieved.

5.2.2 Problem Formulation and Modeling

For an input image I, an NR-IQA model $\mathcal{M}$ is used to estimate the perceptual quality of this image Q_{est}:

$$Q_{est} = \mathcal{M}(I;\theta), \tag{5.1}$$

where θ indicates all of the parameters of this model. Denote the ground truth quality of this image as Q_{gt}, the training objective of this model $\mathcal{M}$ is to find the optimal parameter setting $\hat{\theta}$ so that the quality evaluation loss $\mathcal{L}_q$ between the Q_{est} and Q_{gt} of all test images in the evaluated dataset is in its minimum. Due to our preliminary experiments, we consider the $\ell 1$-norm instead of widely used $\ell 2$-norm

as our $\mathcal{L}_q$:

$$\mathcal{L}_q = \frac{1}{N}\sum_{i=1}^{N} \parallel Q_{est,i} - Q_{gt,i} \parallel_1, \tag{5.2}$$

where the subscript i of $Q_{est,i}$ and $Q_{gt,i}$ represent the estimated quality score and ground truth quality label of i-th image, respectively. Without loss of generality, we ignore this subscript in the following statements for simplification.

To be specific, we divide the pipeline of our end-to-end deep CNN-based NR-IQA model into several stages according to the change in feature dimension. Firstly, we use a deep CNN (DCNN) as the feature extractor to get the raw CNN features f_{raw} directly from the input image I with a size of $h \times w \times 3$. In our implementation, we use one of two commonly used backbone networks, VGG-16 [65] and ResNet-50 [66]. Within these two backbone networks, their fully connected layers are discarded since we need the feature maps with spatial information for further feature fusion. The spatial dimension of these feature maps generated from the last layer of any of these two backbones is $\frac{h}{32} \times \frac{w}{32}$. To ensure the output channels of these feature maps are also the same, an extra $1 \times 1 \times 512$ convolutional layer is added, which can also perform the feature adaption, at the end of the backbone network. We treat this modified network as our feature extractor and use the $DCNN(\cdot)$ to denote this mapping: $\mathbb{R}^{h \times w \times 3} \rightarrow \mathbb{R}^{\frac{h}{32} \times \frac{w}{32} \times 512}$. Therefore, this feature extraction processing can be simply represented by:

$$\textbf{Feature Extraction:} \quad f_{raw} = DCNN(I; \theta_1), \tag{5.3}$$

where θ_1 indicates the weights of the convolutional layers within this DCNN. These raw features f_{raw} can be further processed to become more discriminative feature maps f_m by fusing the saliency information with them as our models, which will be detailed in the next section. As a comparison, we build our baseline model by direct feeding these raw features to the remaining quality evaluation sub-network. It means $f_m = f_{raw}$ in this baseline model. End-to-end CNN-based NR-IQA models usually adopt a multi-layer perceptron (MLP) to regress the image/patch features into a subjective quality score. However, feature maps f_m cannot be directly fed into the MLP. A conversion from feature maps f_m to the image feature vector f_v is required. Here, we empirically choose the global average pooling (GAP) to perform this conversion: $\mathbb{R}^{\frac{h}{32} \times \frac{w}{32} \times 512} \rightarrow \mathbb{R}^{512}$. Our quality evaluation sub-network can be written in the following simultaneous equations, where θ_2 indicates the parameters

inside the MLP.

$$\textbf{Quality Evaluation:} \quad \begin{cases} f_v = GAP(f_m), \\ Q_{est} = MLP(f_v; \theta_2). \end{cases} \tag{5.4}$$

5.2.3 Direct and Multi-task SGDNet

In the feature fusion stage, we propose to incorporate the saliency information into the raw extracted features for getting perceptually-consistent feature maps f_m. Considering that a saliency map is used to measure the visual importance of local regions within its corresponding input image, we implement this property by using element-wise multiplication $\otimes$ to combine the raw features f_{raw} and saliency map S_m together, as shown in Eq. 5.5. As such, this saliency map will be treated as spatial attention prior, where the value of each pixel belongs to [0,1], to re-weight the raw features of the corresponding input image in its spatial domain for a perceptually-consistent feature fusion. Note that the spatial dimension of raw feature maps is not matched with the original size of the generated saliency map, we need to resize the original saliency map with a down-sampling rate of 32 to make this multiplication meaningful.

$$\textbf{Feature Fusion:} \quad f_m = f_{raw} \otimes S_m. \tag{5.5}$$

The roles of these obtained saliency maps in our two variants of SGDNets are different. In our Direct SGDNet, saliency maps are directly used as one of the model inputs for providing additional guidance. By contrast, our Multi-task SGDNet takes these saliency maps as the immediate regression targets by using a shallow CNN (SCNN) as the saliency prediction sub-network to predict them, as shown in Fig. 5.2. In other words, both of these two variants of SGDNets are all using the S_m as the spatial attention prior. The difference between them is that S_m is learned before the feature fusion in the multi-task version or directly provided in the direct one. The processing of our saliency prediction sub-network can be represented by:

$$\textbf{Saliency Prediction:} \quad S_m = SCNN(f_{raw}; \theta_3), \tag{5.6}$$

where θ_3 indicates the parameters of this SCNN. For predicting visual saliency within this SCNN, a saliency prediction loss $\mathcal{L}_s$ should be taken into consideration

to measure the gap between the predicted saliency prior and its corresponding proxy ground truth saliency map. By viewing the saliency map as a kind of probability distribution, as done in Chapter 3, we adopt the total variation distance as this $\mathcal{L}_s$:

$$\mathcal{L}_s(x^p, x^g) = \frac{1}{2}\sum_{i=1}^{N} |\frac{x_i^p}{\sum_{i=1}^{N} x_i^p} - \frac{x_i^g}{\sum_{i=1}^{N} x_i^g}|, \tag{5.7}$$

where $x = (x_1, ..., x_i, ..., x_N)$ is the set of raw saliency response values for either the predicted saliency map (x^p) and the ground truth saliency map (x^g). We have also tried other probability distribution distance metrics, such as Kullback-Leibler divergence, as this saliency prediction loss and observed a similar performance. For jointly optimizing this multi-task SGDNet, an overall loss function should be defined. Here we simply use a linear combination of $\mathcal{L}_q$ and $\mathcal{L}_s$ to represent our optimization target of this SGDNet:

$$\hat{\theta} = \arg\min_{\theta}(\mathcal{L}_q + \alpha\mathcal{L}_s), \tag{5.8}$$

where $\theta = (\theta_1; \theta_2; \theta_3)$ in this model and α is a non-negative parameter to control the relative importance of the saliency prediction sub-task.

5.2.4 Spatial Attention and Channel-wise Attention

As introduced in [9], CNNs extract features by fusing spatial and channel-wise information together, which means there are some inter-dependencies between the channels of CNN features. The channel-wise attention (CA) mechanism is explicitly modeled by their proposed Squeeze-and-Excitation (SE) block to selectively emphasize informative feature channels and suppress less useful ones. Since spatial attention and channel-wise attention are intended to re-calibrate the CNN features in the spatial and channel dimensions separately, we can further extend our model with the CA mechanism for obtaining a more comprehensive feature representation. Here, one SE block [9] is directly plugged into our framework for modeling CA, as depicted in Fig. 5.3. In this case, the previous feature fusion stage should be modified by:

$$\textbf{Feature Fusion with CA:} \quad \begin{cases} f_{CA} = SE(f_{raw}; \theta_4), \\ f_m = f_{CA} \otimes S_m. \end{cases} \tag{5.9}$$

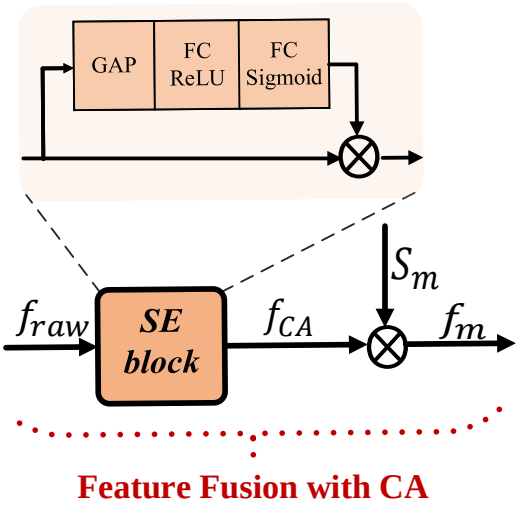


FIGURE 5.3: Illustration of feature fusion with channel-wise attention (CA) in our framework. Within the SE block [9], FC block with an activation function name indicated in the bottom represents the fully-connected layer followed with that specific activation layer.

where θ_4 represents the parameters within this SE block. To avoid unnecessary interactions between the spatial attention and channel-wise attention in our architecture, the raw features from the feature extractor are processed with these two attention mechanisms separately in a parallel manner. Our experimental results show that incorporating CA can improve the performance of our baseline model, but not as good as our proposed spatial attention, which is learned from the saliency prediction sub-network.

5.3 Experiments

5.3.1 IQA Datasets

We perform experiments on two types of image quality datasets covering the synthetic and authentic distortions, respectively. In the former type of IQA datasets, including LIVE [161], CSIQ [162], and TID2013 [11], the distorted images are generated by simulating a single type of synthetically introduced distortion, such as JPEG compression, Gaussian blur, or white noise, with several pre-defined distortion levels on the pristine images. However, massive real-world images suffer from a mixture of diverse and complex authentic distortions that cannot be well-simulated

by the above synthetic distortions. Therefore, new IQA datasets, including CLIVE [163] and KonIQ-10k [7], have been generated to investigate the images with authentic distortions. In CLIVE, images are captured by a wide variety of mobile camera devices under highly diverse conditions. In KonIQ-10k, images are sampled from the massive quantity of Internet images and then filtered to ensure the content diversity and distortion authenticity. Both CLIVE and KonIQ-10k datasets have no reference images, and thus only NR-IQA methods can be used to evaluate them. For better readability, we provide a detailed information summary of the above IQA datasets in Table. 5.1. It should be noted that the subjective score types, subjective score ranges, and image resolutions of these datasets are not unified.

TABLE 5.1: Information summary of IQA datasets

Dataset	Year	# Reference images	# Distorted images	Distortion Type	# Distortion Types	Score type	Score range	Image Resolution
LIVE [161]	2006	29	779	synthetic	5	DMOS	[0,100]	mostly 512×768
CSIQ [162]	2009	30	866	synthetic	6	DMOS	[0,1]	512×512
TID2013 [11]	2013	25	3000	synthetic	24	MOS	[0,9]	384×512
CLIVE [163]	2016	N/A	1162	authentic	N/A	MOS	[0,100]	mostly 500×500
KonIQ-10k [7]	2017	N/A	10073	authentic	N/A	MOS	[1,5]	768×1024

5.3.2 Evaluation Metrics for IQA

Two commonly used evaluation metrics, Spearman rank order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC), are adopted to benchmark the IQA models. Both metrics measure the correlation between a set of objective quality scores Q_{est} estimated/predicted by IQA algorithms and a set of subjective quality scores Q_{gt} provided by subjective experiments.

$$SRCC(Q_{est}, Q_{gt}) = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}, \tag{5.10}$$

$$PLCC(Q_{est}, Q_{gt}) = \frac{cov(Q_{est}, Q_{gt})}{\sigma(Q_{est})\sigma(Q_{gt})}, \tag{5.11}$$

where N is the total number of images for evaluating; d_i is the rank difference between the Q_{est} and Q_{gt} of the i-th image; $cov(\cdot, \cdot)$ and $\sigma(\cdot)$ refer to the covariance and standard deviation, respectively. Generally, SRCC measures the prediction monotonicity, and PLCC measures the prediction accuracy. For both metrics, a

higher value close to 1 indicates the higher performance of a specific IQA method. Some of the methods use a nonlinear function to map their raw model predictions to the MOS scale on the tested dataset for obtaining higher PLCC scores. But it is not adopted in our work since our SGDNet can obtain desirable results on that scale.

5.3.3 Implementation Details

Our SGDNet model is implemented by Keras [139]. The backbone network of SGDNet is VGG-16 or ResNet-50. In our saliency prediction sub-network, the SCNN module is simply implemented by one 1×1 convolutional layer which can achieve the mapping from the raw features to the saliency maps. Moreover, the MLP applied in the remaining quality evaluation sub-network consists of 3 hidden layers with neuron sizes 1024, 1024, and 1. The α used in computing the overall loss is set to 0.25 for highlighting the importance of the quality evaluation task.

During training, the weights in the backbone network are initialized from its Imagenet [36] pre-trained model. The weights of the remaining layers are initialized by the default setting of Keras. All of the models in our experiment are trained with the widely used Adam optimizer with an initial learning rate of 10^{-4}. This learning rate will be scaled down by a factor of 0.1 after every five epochs without validation loss decreasing. For each dataset, except KonIQ-10k, all of the images are resized to the dominant image resolution of this dataset for mini-batch training. Because of the limited GPU memory, the batch size varies according to the resolution of input images. It is worthwhile mentioning that our SGDNet takes only 0.020s for evaluating the image quality of one input image of size 384×512 by using one single GTX 1080 Ti GPU. The source codes of our SGDNet and its pre-trained models are publicly available[2].

[2]https://github.com/ysyscool/SGDNet

TABLE 5.2: Performance comparison on four individual datasets. In each column, the best and second best results are highlighted in **boldface** and ***boldface italic***, respectively.

Methods	LIVE [161]		CSIQ [162]		TID2013 [11]		CLIVE [163]		Weighted Average	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BLIINDS-II [122]	0.912	0.916	0.780	0.832	0.536	0.628	0.463	0.507	0.608	0.673
DIIVINE [121]	0.925	0.923	0.835	0.817	0.549	0.654	0.509	0.558	0.634	0.695
BRISQUE [123]	0.939	0.942	0.775	0.781	0.572	0.651	0.607	0.645	0.659	0.708
IL-NIQE [168]	0.902	0.908	0.821	0.865	0.521	0.648	0.594	0.589	0.631	0.703
CORNIA [124]	0.942	0.943	0.714	0.781	0.549	0.613	0.618	0.662	0.640	0.692
HOSA [118]	0.948	0.949	0.781	0.841	0.688	0.764	0.659	0.678	0.731	0.783
Kang's CNN [46]	0.956	0.956	-	-	-	-	-	-	-	-
BIECON [125]	0.961	0.962	0.825	0.838	0.721	0.765	0.595	0.613	0.743	0.771
DIQaM-NR [133]	0.960	0.972	-	-	***0.835***	***0.855***	0.606	0.601	-	-
WaDIQaM-NR [133]	0.954	0.963	-	-	0.761	0.787	0.671	0.680	-	-
VIDGIQA [127]	***0.969***	***0.973***	-	-	-	-	0.701	-	-	-
ResNet-50 + fine-tuning [166]	0.950	0.954	0.876	***0.905***	0.712	0.756	***0.819***	***0.849***	0.790	0.823
Imagewise CNN [166]	0.963	0.964	0.812	0.791	0.800	0.802	0.663	0.705	0.796	0.803
DIQA [167]	**0.975**	**0.977**	**0.884**	**0.915**	0.825	0.850	0.703	0.704	***0.830***	***0.848***
SGDNet (ours)	***0.969***	0.965	***0.883***	0.903	**0.873**	**0.861**	**0.851**	**0.872**	**0.883**	**0.883**

5.3.4 Performance Comparison

5.3.4.1 Performance on Individual Datasets

In this part, we conduct individual dataset evaluation on four IQA datasets, including LIVE [161], CSIQ [162], TID2013 [11], and CLIVE [163]. Among them, the first three datasets are for synthetic distortions while CLIVE is for authentic distortions. The KonIQ-10k dataset is not adopted in this experiment as it is not used by most of the compared methods. Following the experimental setting of [166], for each individual dataset, we randomly divide it into two subsets according to the reference images, 80% of the data for training and the remaining for test. Then, the corresponding distorted images can be divided into two subsets with non-overlapping image contents. The SRCC and PLCC results of our method are averaged after ten repetitions of this random process.

In Table 5.2, we compare the performance of our ResNet-based SGDNet with six traditional NR-IQA methods, shown in the first six rows, and eight CNN-based NR-IQA methods, starting from the Kang's CNN [46] to the up-to-date DIQA [167]. Although the MEON [45] model is one of the most relevant methods to our SGDNet, it is not included in this performance comparison because it has a quite different experimental setup from the methods listed in this table.

For SRCC and PLCC scores on each dataset, the best and second-best models among these NR-IQA methods are highlighted in **boldface** and ***boldface italic***,

respectively. Moreover, the weighted average results over these four datasets are also reported in the last two columns of Table 5.2. The weight of each database is proportional to the number of distorted images in that database. From this table, we can conclude that our method achieves state-of-the-art performance on both authentically and synthetically distorted IQA datasets.

Besides, we can observe that CNN-based methods are generally superior to traditional methods. Among those existing NR-IQA methods, except for several ResNet-based models, the performance results on the CLIVE dataset is much lower than on the other three synthetically distorted datasets. We interpret this phenomenon as follows. The real-world images with diverse authentic distortions in the CLIVE dataset have a much wider range of image contents which cannot be well handled by those traditional NR-IQA methods without using the deep CNN features. With the help of the ResNet-50 backbone network for feature extraction and the incorporation of saliency information learned by our multi-task framework, our SGDNet largely outperforms all of the compared methods in this authentically distorted IQA dataset. Apart from the CLIVE dataset, our model also achieves competitive performance on the other three synthetically distorted IQA datasets, especially on the TID2013 dataset. Our model is not as good as DIQA [167] on the LIVE and CSIQ datasets. The patch-based training strategy used in DIQA is more suitable for the evaluation of these two relatively small datasets with limited synthetic distortions but is not reliable for the evaluation of the CLIVE dataset.

5.3.4.2 Performance on Individual Distortion Types

Following [45, 167], we train and test our SGDNet on the full TID2013 database to investigate its behaviors on individual distortion types. The detailed experimental setting is the same as the individual dataset evaluation. The SRCC results of this study are reported in Table 5.3. We observe similar results by using PLCC as the evaluation metric. From this table, we can see that SGDNet achieves the best two performing models on most distortion types, as well as the best overall performance. Specifically, most existing methods are not good at several distortion types in TID2013, such as local block-wise distortions, mean intensity shift, contrast change, and change of color saturation. SGDNet performs relatively better on these particular distortions which makes it superior in the overall comparison.

TABLE 5.3: SRCC results of individual distortion types on TID2013 [11]. In each row, the best and second best results on this distortion type are highlighted in **boldface** and ***boldface italic***, respectively.

Distortions\Methods	BLIINDS-II [122]	DIIVINE [121]	BRISQUE [123]	IL-NIQE [168]	CORNIA [124]	HOSA [118]	RankIQA [169]	MEON [45]	DIQA [167]	**SGDNet**
Additive Gaussian noise	0.714	0.756	0.674	**0.924**	0.496	0.833	0.667	0.813	***0.915***	0.877
Additive noise in color components	0.728	0.464	0.550	**0.847**	0.130	0.575	0.620	0.722	0.755	***0.762***
Spatially correlated noise	0.825	0.869	0.804	**0.947**	0.655	0.808	0.821	0.926	0.878	***0.944***
Masked noise	0.358	0.374	0.222	**0.786**	0.373	0.432	0.365	0.728	0.734	***0.756***
High frequency noise	0.852	0.794	0.824	0.908	0.715	0.906	0.760	***0.911***	**0.939**	0.861
Impulse noise	0.664	0.704	0.749	***0.847***	0.647	0.817	0.736	**0.901**	0.843	0.836
Quantization noise	0.780	0.650	0.677	**0.933**	0.632	0.783	0.783	***0.888***	0.858	0.885
Gaussian blur	0.852	0.900	0.855	0.869	0.844	0.903	0.809	0.887	***0.920***	**0.961**
Image denoising	0.754	0.814	0.492	0.846	0.688	**0.873**	0.767	0.797	0.788	***0.852***
JPEG compression	0.808	0.795	0.751	0.901	0.758	**0.903**	0.866	0.850	***0.892***	0.827
JPEG2k compression	0.862	0.804	0.696	**0.930**	0.866	***0.920***	0.878	0.891	0.912	0.897
JPEG transmission errors	0.251	0.514	0.285	0.400	0.587	0.712	0.704	0.746	**0.861**	***0.829***
JPEG2k transmission errors	0.755	***0.892***	0.719	0.708	0.603	0.743	0.701	0.716	0.812	**0.912**
Non-eccentricity pattern noise	0.081	0.215	0.158	-0.173	0.282	0.143	***0.512***	0.116	**0.659**	0.353
Local block-wise distortions	0.371	0.289	0.362	0.000	-0.025	0.330	***0.622***	0.500	0.407	**0.739**
Mean intensity shift	0.159	0.124	0.253	***0.328***	0.194	0.279	0.268	0.177	0.299	**0.643**
Contrast change	-0.082	0.189	0.102	0.080	0.145	0.307	0.613	0.252	***0.687***	**0.811**
Change of color saturation	0.109	0.280	0.200	0.103	-0.006	0.414	0.662	***0.684***	-0.151	**0.749**
Multiplicative Gaussian noise	0.699	0.691	0.587	0.773	0.461	0.711	0.619	0.849	**0.904**	***0.869***
Comfort noise	0.222	0.340	0.211	0.507	0.560	0.537	0.644	0.406	***0.655***	**0.766**
Lossy compression of noisy images	0.451	0.690	0.546	0.911	0.648	0.756	0.800	0.772	**0.930**	***0.914***
Image color quantization with dither	0.815	0.769	0.842	0.822	0.646	0.840	0.779	***0.857***	**0.936**	0.776
Chromatic aberrations	0.568	0.700	0.770	***0.801***	0.672	**0.821**	0.629	0.779	0.756	0.402
Sparse sampling and reconstruction	0.856	0.795	0.764	0.878	0.867	***0.903***	0.859	0.855	**0.909**	0.885
ALL	0.550	0.632	0.572	0.534	0.611	0.707	0.780	0.808	***0.825***	**0.873**

TABLE 5.4: SRCC results in the cross dataset evaluation. In each column, the top result is highlighted in **boldface**.

Methods	CSIQ [162]	TID2013 [11]	CLIVE [163]
BLIINDS-II [122]	0.577	0.393	0.119
DIIVINE [121]	0.590	0.355	**0.465**
BRISQUE [123]	0.548	0.358	0.313
CORNIA [124]	0.649	0.360	0.443
VIDGIQA [127]	0.641	0.415	0.315
DIQaM-NR [133]	0.681	0.392	-
WaDIQaM-NR [133]	0.704	0.462	-
SGDNet (ours)	**0.719**	**0.532**	0.455

Besides, non-eccentricity pattern noise and chromatic aberrations are the disadvantaged subjects of our model. We can further improve SGDNet by investigating these noise patterns.

5.3.4.3 Cross Dataset Evaluation

To test the generalization ability of our method, our proposed model is trained on the entire LIVE dataset and evaluated on the whole TID2013, CSIQ, and CLIVE datasets. Note that the score ranges and score types are not unified in these four datasets as shown in Table 5.1, we choose the settings of CSIQ as our standard in these experiments. Therefore, subjective scores on the other three datasets are linearly scaled to the range of [0,1]. For the MOS values in TID2013 and CLIVE, they are further reversed as $1 - MOS$ to meet this standard. Although our model can process input images with arbitrary sizes, we find that the cross dataset performance of our model achieves the local extrema by resizing all of the training and test images to 384×512. The results of the cross dataset evaluation are reported in Table 5.4 where we can observe that our proposed SGDNet model has a promising generalization ability compared to other methods.

5.3.4.4 Performance on KonIQ-10k dataset

Recently, the authors of KonIQ-10k dataset [7] benchmarked some traditional NR-IQA methods and two recent CNN models, i.e., DeepRN [170] and KonCept512 [7], in their dataset. By following their training/test dataset division, we report our methods and retrained KonCept512 in Table 5.5. Our ResNet-based SGDNet can

TABLE 5.5: Performance comparison on KonIQ-10k test dataset [7]

Method	BLIINDS-II [122]	BRISQUE [123]	CORNIA [124]	HOSA [118]	DeepRN [170]	KonCept512 [7]	SGDNet	SGDNet
Backbone	-	-	-	-	ResNet-101	InceptionResNetV2	ResNet-50	InceptionResNetV2
SRCC	0.585	0.705	0.780	0.805	0.867	0.900	0.899	0.909
PLCC	0.598	0.707	0.808	0.828	0.880	0.920	0.917	0.929

TABLE 5.6: Model ablation analysis on KonIQ-10k dataset [7]

Model Type	Backbone Network		Saliency Information		CA	Performance		
	VGG-16	ResNet-50	saliency input	saliency output		SRCC ↑	PLCC ↑	MAE ↓
Baseline	✓					0.817	0.822	0.2714
		✓				0.808	0.816	0.2532
Baseline + CA	✓				✓	0.827	0.853	0.2623
		✓			✓	0.833	0.851	0.2370
Direct SGDNet	✓		✓			0.843	0.874	0.2017
		✓	✓			0.869	0.890	0.1938
Direct SGDNet + CA	✓		✓		✓	0.851	0.880	0.1963
		✓	✓		✓	0.880	0.899	0.1830
SGDNet	✓			✓		0.870	0.890	0.1881
		✓		✓		0.897	0.917	0.1684
SGDNet + CA	✓			✓	✓	0.878	0.896	0.1846
		✓		✓	✓	0.903	0.920	0.1639

obtain a very close performance towards the KonCept512 model which is equipped with a more powerful InceptionResNetV2 [171] backbone network. By changing our backbone to this powerful one, SGDNet can outperform KonCept512 by a clear margin.

5.3.5 Ablation Study

To evaluate the contribution of each component in our SGDNet, we conduct a series of ablation experiments on the KonIQ-10k dataset [7]. KonIQ-10k is the current largest IQA dataset of authentically distorted images and is suitable for evaluating the performance of different CNN-based NR-IQA models. For this dataset, we randomly pick 8,073 out of the total 10,073 images as the training set and the remaining images as the test set. All of the images and their original saliency maps are resized to 384×512 for accelerating the training process and using a relatively large batch size (20 in this case) for training. The major ablation results are presented in Table 5.6. Apart from the introduced SRCC and PLCC, MAE is also used as the evaluation metric in this study, as it is also the quality evaluation loss that can represent the goodness of fit of these models in this table. Different from the SRCC and PLCC, the lower MAE value indicates a better performance.

5.3.5.1 Influence of the backbone network

In this ablation experiment, we implemented our model with two different backbone networks including VGG-16 [65] and ResNet-50 [66]. It is widely accepted that ResNet-50 is more powerful than VGG-16, as it can be trained very deeply for more comprehensive feature extraction with the help of residual learning. In our case, we compare the models listed in Table 5.6 with different backbone networks. In our baseline models, some VGG-based models have better performance concerning SRCC or PLCC. The reason is that SRCC and PLCC scores may not be always consistent with our quality evaluation loss. In our proposed saliency-guided models, including Direct SGDNet and (Multi-task) SGDNet, the ResNet-based models perform better than their VGG-based versions on all of the three evaluation metrics.

5.3.5.2 Effectiveness of saliency information

We incorporate the saliency information in two different ways. In our Direct SGDNet, the obtained saliency maps are served as one of the model inputs. Extracted features from the backbone network are further fused with these fixed saliency maps. As a comparison, our SGDNet is built on an end-to-end multi-task framework and these saliency maps are the target outputs of our proposed saliency prediction sub-network. The predicted saliency maps of our sub-network have the same function as the fixed ones for providing the spatial attention priors. From Table 5.6, we can observe that the quality evaluation performance indeed boosted by incorporating the saliency information. Moreover, by comparing the Direct SGDNet and SGDNet with the same configuration of the backbone and CA, we can find the latter model is always better. It means that our end-to-end multi-task learning framework can further improve the boosted performance achieved by simply using the saliency information as the additional model input.

Fig. 5.4 provides some examples to compare those targets and learned saliency maps. Saliency maps generated by our DINet[8] have some undesirable responses on the image centers because this saliency model is trained on the saliency prediction datasets where center-bias priors are implicitly learned. Our saliency prediction sub-network can skip this trap by jointly training with the quality evaluation sub-network while the center regions in these input images are not appealing to the

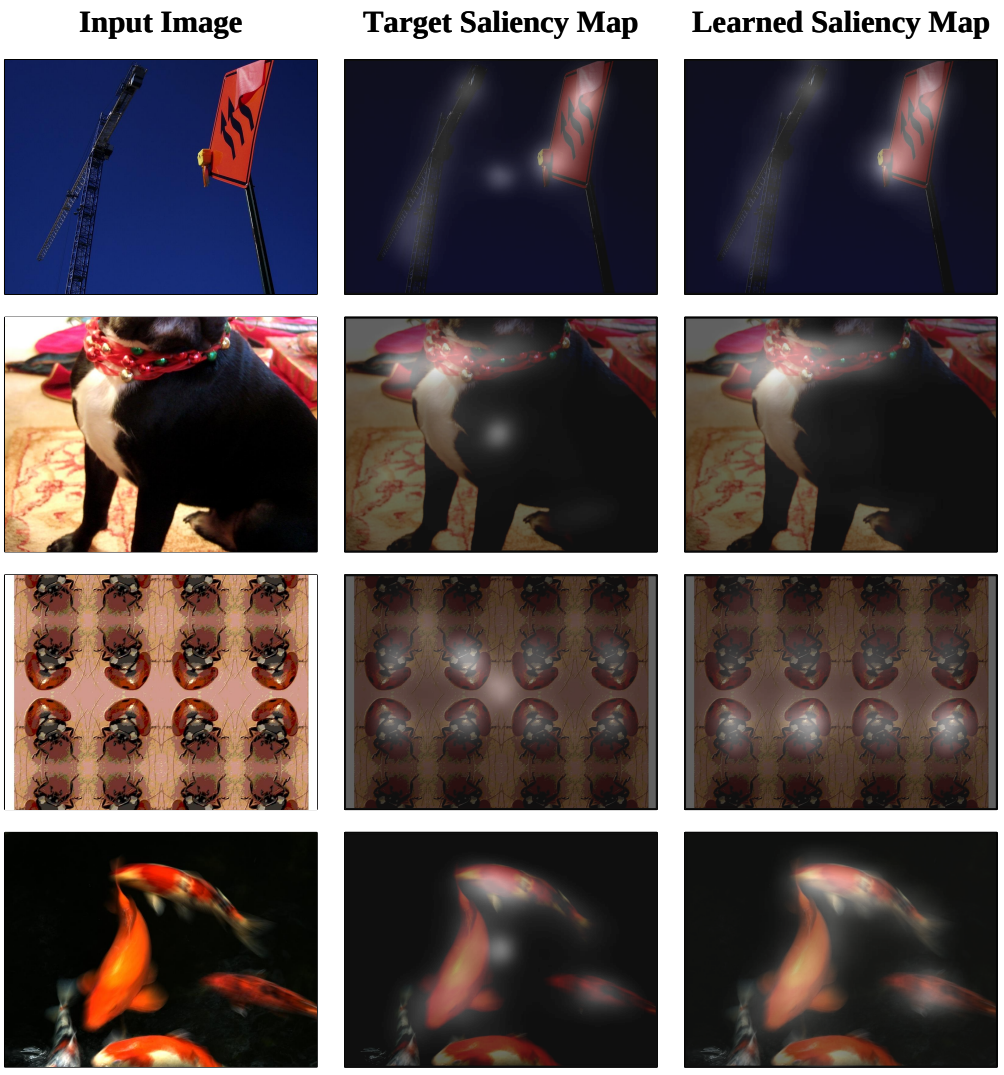


FIGURE 5.4: Examples of input images, target saliency maps generated by [8], and predicted saliency maps by our proposed saliency prediction sub-network.

later quality evaluation sub-network. Therefore, we conclude that the saliency maps learned in SGDNet are more adaptive and suitable than those fixed ones.

5.3.5.3 Influence of channel-wise attention

Besides modeling the spatial attention by using the saliency information, our method can be further extended by incorporating the channel-wise attention (CA) module, as described in Section 5.2.4. From Table 5.6, in the baseline models, the introduced CA module [9] can improve the performance by a large margin (more than 3.7% on PLCC and 1.2% on SRCC). However, we also observe that the spatial attention prior is more powerful than CA by comparing the models in the

TABLE 5.7: PLCC results of saliency models on KonIQ-10k IQA dataset [7] with our ResNet-based SGDNet and MIT300 saliency dataset [12] with themselves.

Saliency models	KonIQ-10k	MIT300
DINet [8]	0.917	0.79
SAM [67]	0.914	0.78
Our SOD model [107]	0.912	-
SalGAN [70]	0.911	0.73
DVA [38]	0.903	0.68
(no saliency model)	0.816	-

second and third types in Table 5.6. It can be explained that incorporating saliency information can lead to a more perceptually-consistent feature fusion while incorporating CA can only explicitly modeling the inter-dependencies between feature channels which is not much related to the quality evaluation task. Furthermore, by fusing the spatial and channel-wise features together, our extended SGDNet (SGDNet + CA) can achieve slightly higher results when comparing with the original model.

5.3.6 Influence of the Saliency Model

Intuitively, the performance of our SGDNet is also affected by the teacher saliency model, which produces the proxy saliency maps for training our saliency prediction sub-network. To explore the influence of this fact, we test other three most relevant saliency models, including SAM [67], SalGAN [70], and DVA [38], to replace the currently used DINet [8]. Since PLCC is the common evaluation metric on both quality evaluation and saliency prediction, we report the PLCC results of different saliency models with our ResNet-based SGDNet on KonIQ-10k IQA dataset [7] and their own performance on MIT300 saliency benchmark dataset [12] in Table 5.7. By comparing the results in the second column, we can find that the effectiveness of saliency information is still validated by using different saliency models. Moreover, the performance of our SGDNet has a positive correlation with the performance of the used saliency models in the MIT300 saliency benchmark dataset [12]. Specifically, using SAM, SalGAN, or DVA to replace DINet will lead to performance degradations by 0.3%, 0.6%, 1.5%, respectively. It can be foreseen that our SGDNet can be further improved with the advent of more powerful saliency prediction models.

Besides the visual saliency prediction models, we also try to use our SOD model [107] to generate such proxy saliency maps for the saliency prediction sub-task in our SGDNet. The PLCC result of this approach is also presented in Table 5.7. We can find that although our SOD model outperforms our DINet in many SOD benchmark datasets, as shown in Chapter 4, DINet is more suitable than this SOD model in evaluating image quality. SOD model can provide much more complete and accurate saliency priors for the salient objects. However, not all of these regions are perceptually-consistent with image quality. In other words, the saliency maps provided by the SOD model contain several regions that are redundant in evaluating image quality. Therefore, for perception-aware applications, visual saliency based on human fixations can better represent the perceptually-consistent information.

5.4 Summary

In this work, we have proposed a novel saliency-guided deep neural network (SGDNet) for no-reference image quality assessment (NR-IQA). The whole model is built on an end-to-end multi-task learning framework where two sub-tasks, including visual saliency prediction and image quality evaluation, are jointly optimized and have certain dependencies on each other. The effectiveness of incorporating learnable or fixed saliency information and our multi-task framework for CNN-based NR-IQA have been validated by a series of ablation studies. Moreover, our method overcomes the inability of existing multi-task CNN-based NR-IQA methods on evaluating the perceptual qualities of real-world images with authentic distortions. Experimental results on both authentically and synthetically distorted IQA datasets have demonstrated the outstanding performance of our model with respect to other relevant NR-IQA methods. In the future, we will consider a new spatial attention mechanism that can learn the spatial attention maps from the image itself instead of using a proxy saliency map as supervision.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we have presented our research work on building new deep learning-based computational models for visual saliency computation and quality evaluation. In the literature, there are two types of saliency models for different tasks, including visual saliency prediction and salient object detection (SOD). Currently, the popular deep learning-based saliency models for these two tasks are all based on fully convolutional network architecture, which means that both types of saliency models can be applied interchangeably. However, as shown in Section 4.3.6, we find that our fixation prediction model applied in SOD is not good as some existing works which are specifically designed for SOD. The architectures of the saliency models for fixation prediction are much simpler than those for SOD with multiple feature aggregation modules and powerful refinement modules. It is reasonable since SOD is more complex toward the visual saliency prediction task as the labeled saliency maps in SOD is more dense and structured. While for a perception-aware application like image quality assessment, saliency prediction models are more useful than those SOD approaches in providing better perceptually-consistent information. For each individual research work, our main contributions are summarized as follows.

In Chapter 3, we have proposed a dilated inception network for visual saliency prediction. It captures diverse multi-scale contextual features effectively with very limited extra parameters by the proposed dilated inception module. The whole model works in a fully convolutional encoder-decoder architecture, which

is trained end-to-end and lightweight for time-efficiency. Besides, a set of linear normalization-based probability distribution distance metrics are proposed as loss functions to optimize our DINet. As such, the saliency prediction task is formulated as a probability distribution prediction problem, consequently leading to an extra performance gain. Experimental results on several challenging saliency benchmark datasets have demonstrated that our DINet with proposed loss functions can achieve state-of-the-art performance with shorter inference time.

In Chapter 4, we have proposed a simple yet effective progressive self-guided loss to facilitate deep learning-based salient object detection. The proposed progressive self-guided loss imitates a morphological closing operation on the model predictions for creating progressive and auxiliary training supervisions epoch-wisely to step-wisely guide the training process. We have demonstrated that this new loss function can guide the salient object detection model to highlight more complete salient objects step-by-step and meanwhile help to uncover the spatial dependencies of the salient object pixels in a region growing manner. Experimental results on six widely used SOD benchmark datasets have shown that our loss function not only advances the performance of existing deep learning-based models without architecture modification but also helps our proposed framework to achieve state-of-the-art performance.

In Chapter 5, we have proposed a novel saliency-guided deep neural network for no-reference image quality assessment. The whole model is built on an end-to-end multi-task learning framework where two sub-tasks, including visual saliency prediction and image quality evaluation, are jointly optimized and have certain dependencies on each other. The effectiveness of incorporating saliency information and our multi-task framework has been validated by a series of ablation studies. Moreover, our method overcomes the inability of existing multi-task CNN-based methods on evaluating the perceptual qualities of real-world images with authentic distortions. Experimental results on both authentically and synthetically distorted image quality assessment datasets have demonstrated the outstanding performance of our model concerning other relevant methods.

6.2 Future Work

This thesis focuses on building new deep learning-based computational models for basic visual saliency computation and saliency-guided image quality assessment. There are many interesting applications of visual saliency waited to be investigated. Besides, the possible extensions of visual saliency computation and quality evaluation are also not explored in this thesis. Based on our current works and the tracking of the latest works in the related areas, we believe the following research directions can further extend this thesis.

- **Visual saliency in 360° video.** The rapid development of virtual reality (VR) has been witnessed in recent years. 360° video, which is an essential type of VR content, is also growing rapidly and available in our daily life [172]. It is essential to build some new visual saliency models to understand and analyze the behavior of a human in VR scenes. Accurate saliency prediction can enable bandwidth-efficient 360° video streaming, where the client only downloads the salient video portions that the user is likely to view in high quality while the remaining portions are ignored or fetched in low-quality [173].

- **Visual saliency in 3D point cloud.** The 3D point cloud is one of the most important 3D data representations for newly-emerged computer vision applications, such as robot perception, augmented reality, and identity recognition [174, 175]. Visual saliency can be introduced and applied to the 3D scenarios for identifying the visually important or machine-perception-aware points from a specific point cloud. Such information may be useful in the point cloud simplification and compression.

- **Driver attention prediction.** Robust driver attention prediction is challenging yet essential for safe and autonomous driving. Recently, several large-scale datasets [176, 177] for this problem have emerged. Visual saliency models can be applied to this task for predicting driver attention under various occasions, weather, and light conditions.

- **Instance-level salient object segmentation.** In the common salient object detection task, we only segment the salient objects from the background and treat this task as a binary segmentation problem. A more challenging

and more useful task is to perform the instance-level segmentation for salient objects [178]. The next generation of salient object detection methods needs to perform more detailed parsing within detected salient regions to achieve this goal. Moreover, our proposed progressive self-guided loss should be modified to properly separate overlapping objects in this challenging task.

- **Perception-aware image restoration.** Existing image quality methods only provide a single quality value for a given image. However, a large number of image processing and computer graphics applications require localized information on the image distortions [179]. One possible direction is to extend the current image quality methods by outputting the locations of distorted regions additionally and providing some modification suggestions for the following perceptual image restoration.

Publications

Journal Articles

- **Sheng Yang**, Guosheng Lin, Qiuping Jiang, and Weisi Lin. "A dilated inception network for visual saliency prediction." *IEEE Transactions on Multimedia* (2019).
- **Sheng Yang**, Weisi Lin, Guosheng Lin, Zichuan Liu, and Qiuping Jiang. "Progressive Self-Guided Loss for Salient Object Detection." In Submission, 2020.
- Qiuping Jiang, Zhenyu Peng, **Sheng Yang**, and Feng Shao. "Authentically Distorted Image Quality Assessment by Learning From Empirical Score Distributions." *IEEE Signal Processing Letters* 26, no. 12 (2019): 1867-1871.

Conference Proceedings

- **Sheng Yang**, Qiuping Jiang, Weisi Lin, and Yongtao Wang. "SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment." In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1383-1391. 2019.
- Jingwen Hou, **Sheng Yang**, and Weisi Lin. "Object-level Attention for Aesthetic Rating Distribution Prediction." In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 816-824. 2020.
- **Sheng Yang** and Weisi Lin. "Predicting visual saliency via a dilated inception module-based model." In *International Workshop on Advanced Image Technology (IWAIT) 2019*, vol. 11049, p. 110491D. International Society for Optics and Photonics, 2019.

- Zichuan Liu, Guosheng Lin, **Sheng Yang**, Fayao Liu, Weisi Lin, and Wang Ling Goh. "Towards robust curve text detection with conditional spatial expansion." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7269-7278. 2019.

- Zichuan Liu, Guosheng Lin, **Sheng Yang**, Jiashi Feng, Weisi Lin, and Wang Ling Goh. "Learning Markov Clustering Networks for Scene Text Detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6936-6944. 2018.

Bibliography

[1] 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. xvii, 3, 13, 14, 17, 30, 35, 36

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. xvii, 14, 17, 30, 36, 38

[3] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. xvii, xix, 14, 21, 28, 30, 38, 42, 44, 46, 48, 51, 54, 55, 57, 84

[4] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009. xvii, xix, 12, 30, 42, 44, 55, 56

[5] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3926, 2019. xvii, 21, 22, 24, 60, 61, 70, 72

[6] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8779–8788, 2019. xvii, 21, 23, 60, 61, 64, 70, 72, 78

[7] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. xviii, xx, 82, 83, 90, 95, 96, 99

[8] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia*, 2019. xviii, 15, 23, 28, 29, 53, 61, 79, 82, 83, 84, 97, 98, 99

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. xviii, 88, 89, 98

[10] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. xix, 30, 43, 45, 56

[11] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. xx, 82, 89, 90, 92, 94, 95

[12] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. xx, 11, 99

[13] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013. 1

[14] Shih-Syun Lin, I-Cheng Yeh, Chao-Hung Lin, and Tong-Yee Lee. Patch-based image warping for content-aware retargeting. *IEEE Transactions on Multimedia*, 15(2):359–368, 2013. 1, 29

[15] Shengxi Li, Mai Xu, Yun Ren, and Zulin Wang. Closed-form optimization on saliency-guided image compression for hevc-msp. *IEEE Transactions on Multimedia*, 2017. 1, 29

[16] Hadi Hadizadeh and Ivan V Bajic. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2014. 1, 29

[17] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia*, 18(6):1098–1110, 2016. 1, 27, 29, 82

[18] Qiuping Jiang, Feng Shao, Weisi Lin, Ke Gu, Gangyi Jiang, and Huifang Sun. Optimizing multistage discriminative dictionaries for blind image quality assessment. *IEEE Transactions on Multimedia*, 20(8), 2018. 24, 25, 81

[19] Haksub Kim and Sanghoon Lee. Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort. *IEEE Transactions on Multimedia*, 17(12):2198–2209, 2015. 1, 29

[20] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013. 1, 29

[21] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. 1, 29

[22] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5): 769–779, 2013. 1, 59

[23] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*, 2014.

[24] Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, and Wang Ling Goh. Learning markov clustering networks for scene text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6944, 2018.

[25] Dingwen Zhang, Deyu Meng, Long Zhao, and Junwei Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. *arXiv preprint arXiv:1703.01290*, 2017. 1, 59

[26] Ming-Ming Cheng, Fang-Lue Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Repfinder: finding approximately repeated scene elements for image editing. *ACM Transactions on Graphics (TOG)*, 29(4):1–8, 2010. 1, 59

[27] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. Saliency for image manipulation. *The Visual Computer*, 29(5):381–392, 2013. 1, 59

[28] Hyemin Lee and Daijin Kim. Salient region-based online object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1170–1177. IEEE, 2018. 1, 59

[29] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2016. 1, 59

[30] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1, 18, 21

[31] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 1, 18, 59

[32] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1711–1720, 2018. 2, 21, 22

[33] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. Inferring salient objects from human fixations. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[34] Ali Borji. Saliency prediction in the deep learning era: Successes, limitations, and future challenges. *arXiv preprint arXiv:1810.03716*, 2018. 3

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 13

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 3, 13, 70, 91

[37] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1155–1162. IEEE, 2013. 3, 14, 69, 70, 72, 75, 77, 78

[38] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *arXiv preprint arXiv:1705.02544*, 2017. 3, 13, 14, 15, 16, 30, 54, 56, 99

[39] Srinivas SS Kruthiventi, Kumar Ayush, and Radhakrishnan Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 2017. 13, 14, 15, 17, 29, 43, 49, 54, 56

[40] Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *arXiv preprint arXiv:1610.01708*, 2016. 3, 13, 14, 15, 29, 30, 33, 43, 45, 49, 50, 51, 54, 56, 57

[41] Kai Zhao, Shanghua Gao, Wenguan Wang, and Ming-Ming Cheng. Optimizing the f-measure for threshold-free salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8849–8857, 2019. 4, 23, 60

[42] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sébastien Ourselin, and Tom Vercauteren. Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks. In *International MICCAI Brainlesion Workshop*, pages 64–76. Springer, 2017. 65

[43] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 4, 21, 23, 24, 60, 64, 72

[44] Le Kang, Peng Ye, Yi Li, and David Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *2015 IEEE international conference on image processing (ICIP)*, pages 2791–2795. IEEE, 2015. 4, 26

[45] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2018. 4, 26, 28, 82, 92, 93, 94

[46] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 25, 26, 92

[47] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018. 4

[48] Jianming Zhang. *Visual saliency computation for image analysis*. PhD thesis, Boston University, 2016. 10, 29

[49] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 10, 18, 29, 56

[50] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007. 10, 12, 29, 56

[51] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013. 10, 11, 12, 29, 56

[52] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006. 11

[53] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007. 11

[54] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in neural information processing systems*, pages 681–688, 2009. 11

[55] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 11

[56] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on*, pages 1–8. IEEE, 2008. 11

[57] Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):996–1010, 2012. 11

[58] Wolf Kienzle, Felix A Wichmann, Matthias O Franz, and Bernhard Schölkopf. A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems*, pages 689–696, 2007. 12

[59] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009.

[60] Hae Jong Seo and Peyman Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 45–52. IEEE, 2009. 12

[61] Chengyao Shen, Mingli Song, and Qi Zhao. Learning high-level concepts by training a deep network on eye fixations. In *NIPS Deep Learning and Unsup Feat Learn Workshop*, volume 2, 2012. 12

[62] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014. 12, 56

[63] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. In *3rd International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1411.1045`. 13, 16

[64] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015. 13, 15, 16, 49, 56

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1409.1556`. 13, 86, 97

[66] 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. 13, 33, 69, 86, 97

[67] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*, 2016. 13, 14, 15, 29, 33, 43, 45, 50, 51, 54, 56, 99

[68] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 13, 20, 59

[69] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017. 13, 15, 54, 56

[70] Junting Pan, Cristian Canton-Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. 2017. URL `http://arxiv.org/abs/1701.01081`. 13, 15, 54, 56, 99

[71] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3488–3493. IEEE, 2016. 13, 15, 16, 54, 56

[72] Samuel Dodge and Lina Karam. Visual saliency prediction using a mixture of deep neural networks. *arXiv preprint arXiv:1702.00372*, 2017. 13, 15, 30, 54

[73] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *4th International Conference on Learning Representations*, 2016. URL `http://arxiv.org/abs/1511.07122`. 14

[74] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 14, 31

[75] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2016. 14, 15, 30, 40, 45, 54, 56

[76] Sen He, Ali Borji, Yang Mi, and Nicolas Pugeault. What catches the eye? visualizing and understanding deep saliency models. 2018. URL `http://arxiv.org/abs/1803.05753`. 14

[77] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 362–370, 2015. 16

[78] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 16

[79] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2376–2383. IEEE, 2010. 18

[80] Meng Wang, Janusz Konrad, Prakash Ishwar, Kevin Jing, and Henry Rowley. Image saliency: From intrinsic to extrinsic context. In *CVPR 2011*, pages 417–424. IEEE, 2011. 18

[81] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014. 18, 21, 59

[82] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, Nanning Zheng, and Shipeng Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, volume 6, page 9, 2011. 18

[83] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013. 19, 59

[84] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on circuits and Systems for Video Technology*, 29(10):2941–2959, 2018. 19

[85] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019. 19, 21, 22, 24, 59, 60, 72

[86] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 19, 20, 21, 69, 70, 72, 75, 77, 78

[87] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015. 19, 21

[88] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *International journal of computer vision*, 115(3):330–344, 2015. 20

[89] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016. 20, 21

[90] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 21, 59, 69, 70, 72, 75, 77, 78

[91] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. 20, 21, 72

[92] Srinivas SS Kruthiventi, Vennela Gudisa, Jaley H Dholakiya, and R Venkatesh Babu. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 20, 21, 54

[93] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision*, pages 825–841. Springer, 2016. 21

[94] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 21, 70

[95] Youbao Tang and Xiangqian Wu. Saliency detection via combining region-level and pixel-level predictions with cnns. In *European Conference on Computer Vision*, pages 809–825. Springer, 2016. 20, 21

[96] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on computer vision*, pages 212–221, 2017. 21, 72

[97] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017. 21, 22, 72, 75

[98] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 6609–6617, 2017. 21, 22, 72

[99] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 21, 22, 23, 64, 72, 75

[100] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4019–4028, 2017. 21, 22, 72

[101] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. 21, 69, 70, 72, 75, 77, 78

[102] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1741–1750, 2018. 21

[103] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3135, 2018. 21, 70, 72

[104] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018. 20, 21, 70, 72

[105] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018. 21, 22, 72

[106] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019. 21, 22, 23, 60, 64, 70, 72, 78

[107] Sheng Yang, Weisi Lin, Guosheng Lin, Zichuan Liu, and Qiuping Jiang. Progressively guided loss for salient object detection. In *Submission*, 2020. 21, 23, 24, 59, 99, 100

[108] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 20, 22

[109] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 22, 61

[110] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 24, 81

[111] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[112] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012.

[113] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 24, 27, 81, 82

[114] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, 21(1):41–52, 2012. 24, 81

[115] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. Reduced-reference image quality assessment with visual information fidelity. *IEEE Transactions on Multimedia*, 15(7):1700–1705, 2013.

[116] Xiongkuo Min, Ke Gu, Guangtao Zhai, Menghan Hu, and Xiaokang Yang. Saliency-induced reduced-reference quality index for natural scene and screen content images. *Signal Processing*, 145:127–136, 2018. 24, 81

[117] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, 2015. 24, 81, 82

[118] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016. 25, 92, 94, 96

[119] Dingquan Li, Tingting Jiang, and Ming Jiang. Exploiting high-level semantics for no-reference image quality assessment of realistic blur images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 378–386. ACM, 2017.

[120] Qiuping Jiang, Feng Shao, Wei Gao, Zhuo Chen, Gangyi Jiang, and Yo-Sung Ho. Unified no-reference quality assessment of singly and multiply distorted stereoscopic images. *IEEE Transactions on Image Processing*, 28(4):1866–1881, 2018. 24, 81

[121] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 25, 82, 92, 94, 95

[122] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 25, 92, 94, 95, 96

[123] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 25, 82, 92, 94, 95, 96

[124] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105. IEEE, 2012. 25, 92, 94, 95, 96

[125] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2017. 25, 92

[126] Cuong T Vu, Eric C Larson, and Damon M Chandler. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. In *2008 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 73–76. IEEE, 2008. 27, 82

[127] Jingwei Guan, Shuai Yi, Xingyu Zeng, Wai-Kuen Cham, and Xiaogang Wang. Visual importance and distortion guided deep image quality assessment framework. *IEEE Transactions on Multimedia*, 19(11):2505–2520, 2017. 27, 28, 82, 83, 92, 95

[128] Hani Alers, Judith A Redi, Hantao Liu, and Ingrid Heynderickx. Studying the effect of optimizing image quality in salient regions at the expense of background content. *Journal of Electronic Imaging*, 22(4):043012, 2013. 27

[129] Wei Zhang and Hantao Liu. Toward a reliable collection of eye-tracking data for image quality research: challenges, solutions, and applications. *IEEE Transactions on Image Processing*, 26(5):2424–2437, 2017. 27, 84

[130] Wei Zhang and Hantao Liu. Learning picture quality from visual distraction: Psychophysical studies and computational models. *Neurocomputing*, 247: 183–191, 2017. 27

[131] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011. 27

[132] Wei Zhang, Ali Borji, Zhou Wang, Patrick Le Callet, and Hantao Liu. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE transactions on neural networks and learning systems*, 27(6):1266–1278, 2015. 27

[133] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018. 27, 28, 83, 92, 95

[134] Yuan Gao, Miaojing Shi, Dacheng Tao, and Chao Xu. Database saliency for fast image retrieval. *IEEE Transactions on Multimedia*, 17(3):359–369, 2015. 29

[135] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003. URL `http://distill.pub/2016/deconv-checkerboard`. 34

[136] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 42

[137] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016. 43

[138] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 1153–1160, 2013. 43

[139] François Chollet et al. Keras. `https://github.com/keras-team/keras`, 2015. 44, 91

[140] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6980`. 44

[141] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in Neural Information Processing Systems*, pages 4467–4475, 2017. 52

[142] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 53

[143] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016. 54, 56

[144] Nian Liu, Junwei Han, Tianming Liu, and Xuelong Li. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 2016. 56

[145] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron C. Courville, and Yoshua Bengio. Renet: A recurrent neural network based alternative to convolutional networks. 2015. URL `http://arxiv.org/abs/1505.00393`. 57

[146] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 57

[147] Junfeng He, Jinyuan Feng, Xianglong Liu, Tao Cheng, Tai-Hsu Lin, Hyunjin Chung, and Shih-Fu Chang. Mobile product search with bag of hash bits and boundary reranking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3005–3012. IEEE, 2012. 59

[148] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 61

[149] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987. 65

[150] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018. 68

[151] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014. 69, 70, 72, 75, 77, 78

[152] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 49–56. IEEE, 2010. 69, 70, 72, 73, 75, 76, 77, 78

[153] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 70

[154] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 70

[155] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1383–1391, 2019. 81

[156] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1709, 2016. 81

[157] Leida Li, Ya Yan, Yuming Fang, Shiqi Wang, Lu Tang, and Jiansheng Qian. Perceptual quality evaluation for image defocus deblurring. *Signal Processing: Image Communication*, 48:81–91, 2016. 81

[158] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018. 81

[159] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 81

[160] Qiaohong Li, Weisi Lin, Jingtao Xu, and Yuming Fang. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia*, 18(12):2457–2469, 2016. 82

[161] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 82, 89, 90, 92

[162] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010. 82, 89, 90, 92, 95

[163] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016. 82, 90, 92, 95

[164] Hantao Liu and Ingrid Heynderickx. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *2009 16th IEEE international conference on image processing (ICIP)*, pages 3097–3100. IEEE, 2009. 84

[165] Judith Redi, Hantao Liu, Rodolfo Zunino, and Ingrid Heynderickx. Interactions of visual attention and quality perception. In *Human Vision and Electronic Imaging XVI*, volume 7865, page 78650S. International Society for Optics and Photonics, 2011. 84

[166] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017. 92

[167] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. Deep cnn-based blind image quality predictor. *IEEE transactions on neural networks and learning systems*, (99):1–14, 2018. 92, 93, 94

[168] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24 (8):2579–2591, 2015. 92, 94

[169] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1040–1049, 2017. 94

[170] Domonkos Varga, Dietmar Saupe, and Tamás Szirányi. Deeprn: A content preserving deep architecture for blind image quality assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 95, 96

[171] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 96

[172] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020. 103

[173] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1190–1198, 2018. 103

[174] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Point-cloud saliency maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1598–1606, 2019. 103

[175] Xiaoying Ding, Weisi Lin, Zhenzhong Chen, and Xinfeng Zhang. Point cloud saliency detection by local and global feature fusion. *IEEE Transactions on Image Processing*, 28(11):5379–5393, 2019. 103

[176] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer, 2018. 103

[177] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018. 103

[178] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2386–2395, 2017. 104

[179] Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K Mantiuk. Dataset and metrics for predicting local visible differences. *ACM Transactions on Graphics (TOG)*, 37(5):1–14, 2018. 104