Deep Texture-Depth-Based Attention for Face Recognition on IoT Devices

Yuxin Lin, Wenye Liu, and Chip-Hong Chang
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
Email: liny0090@e.ntu.edu.sg, wliu015@e.ntu.edu.sg, echchang@ntu.edu.sg

***Abstract*—Traditional face recognition systems use RGB images as input for feature extraction and classification. However, conventional methods based on color images experience non-trivial accuracy drop under several challenging conditions like occlusion, pose variation and facial expression changes. With the gradually decreasing cost of smart sensors, RGB-Depth(D) images captured using low-cost sensors are used to provide complementary features to RGB images. Both the extracted Local Binary Pattern (LBP) features and depth map contain additional discriminative information that can guide the face recognition model to focus on the important parts of the input image. In this paper, we propose a novel end-to-end network that combines both texture and depth features for automatic attention-based face recognition. The experiment results demonstrate that the proposed method has improved recognition accuracy under diverse variations. Our proposed face recognition model has been implemented on the NVIDIA Jetson Nano device to evaluate its performance with compact feature extractors used on different branches of the model. The results show that our method can improve the FPS of face recognition on an edge-coming device from 1.6 to 3.8 with $<1\%$ accuracy degradation.**

***Keywords*—Face Recognition, RGB-Depth, Local Binary Pattern, Attention Fusion**

I. Introduction

With the widespread applications of smart sensors in the industrial Internet of Things (IoT), face recognition has become a main security measure [1] across many applications and industrial verticals from the common usage of user authentication and access control to automatic media content indexing and categorization and human trafficking victims identification. Face recognition is a computer vision task of detecting and identifying a person in a photo or video frame. A variety of machine learning-based face recognition algorithms have been proposed. However, due to various factors like covariate changes, image quality, occlusions, and pose variations, the accuracy of existing RGB-based models can still be improved. There is a growing demand in on-device AI to reduce data transport and latency to enable near real-time decision-making using data generated from smart sensors on or near the endpoint devices.

The process of face recognition consists of three main steps: i) face detection in which human faces are first distinguished from other objects; ii) preprocessing and feature extraction; and iii) identification. Conventional RGB images have been extensively utilized in face recognition since this kind of image is easy to acquire through low-cost sensors. However, when the images contain covariates, such as pose shifts and illumination variations, the accuracy of the RGB-based recognition system will drop significantly. In contrast, RGB-D images have demonstrated their potential to achieve a better recognition performance than RGB images [2], as the depth channel of RGB-D images provides additional discriminative information. With the gradual decrease in the price of depth sensors such as Microsoft Azure Kinect and Intel RealSense Camera, adopting RGB-D images for face recognition has become a viable option [3]. The small footprint of depth sensors allows them to be easily integrated into IoT edge for real-time on-device inference.

Among different deep learning architectures, attention mechanism is effective in understanding images in various computer vision tasks. By learning an attention map that includes weights corresponding to the feature map, it allows a model to focus selectively on the most important regions of an image. For face recognition, the effective features mainly reside at some key parts of the face, such as the eyes, nose, and mouth. The attention approach can highlight these salient features to improve recognition accuracy.

In this paper, our goal is to develop an end-to-end attention-based multi-modal face recognition framework for the edge. We proposed a novel LBP-depth-guided attention model to adaptively focus on the salient features of an input image for RGB-D face recognition. Both the depth images and LBP images derived from RGB images are utilized to generate the attention map for inputs captured under challenging scenarios such as pose variations, occlusion, illumination changes, and facial expression variations. We conducted extensive experiments on Lock3DFace [4] dataset and IIIT-D [5], [6] database to study the effects of different feature-level fusion strategies for attention refinement and different feature extractor combinations. To ensure that the network can still run smoothly on IoT devices, we adopt lightweight feature extractors which reduce the proposed network complexity while keeping the accuracy degradation over the state-of-the-art models to within $\sim 1\%$. Implementation on edge computing devices like NVIDIA Jetson Nano platform is evaluated.

II. Background and Related Works

Traditional two-step machine learning approaches for face recognition rely on appearance-based feature extractors, such as Histogram of Gradients (HoG), Local Binary Patterns (LBP) [7], +and Local Directional Patterns (LDP). These features are extracted based on the statistical relations of a pixel with its neighboring pixels. The extracted features are then fed to a classifier (SVM, Random Forest, Linear Classifier, *etc*.) for identification. These approaches [4], [6], [8] show excellent performance in well-controlled laboratory environments but fall short of expectations when evaluated on more sophisticated and challenging datasets.

In recent years, deep learning approaches to face recognition problems have received extensive attention due to their extraordinary performance in image classification tasks. Convolutional neural networks (CNNs) are widely used in image classification tasks since the weights of the filters in each layer can automatically extract and learn simple low-level to complex high-level features. As such, current state-of-the-art RGB-D face recognition models are mainly designed using CNNs such as VGG [9] and ResNet [10] as backbones. In multi-modal face recognition, additional modalities, such as depth modality, are often combined with RGB to increase the robustness to covariates like expression and pose. To utilize the complementary discriminative information from different modalities, strategies like feature-level fusion and feature pooling are often used. Feature-level fusion fuses the extracted features from two modalities for joint-representation by element-wise addition, multiplication, or concatenation, *i.e*.

$$F_{\text{fusion}} = W_\alpha^T F_a + W_\beta^T F_\beta \tag{1}$$

or

$$F_{\text{fusion}} = W_\alpha^T F_a \circ W_\alpha^T F_\beta. \tag{2}$$

Feature pooling is used to explore the interactions between two modalities. Other than bilinear pooling [11] which uses computationally expensive quadratic expansions, dot product approach [12] has also been studied, *i.e*.

$$F_{\text{pooled}} = \left(W_\alpha^T F_\alpha\right) \odot \left(W_\beta^T F_\beta\right), \tag{3}$$

where W and F are the trainable weights and features, respectively.

One of the earliest attempts to utilize deep learning for RGB-D face recognition was made by Chowdhury *et al*. [13]. A reconstruction mapping from RGB to depth representation was learned and used for data transformation from the input RGB images. The reconstructed depth was then fed into the fully connected (FC) layer of the neural network for identification. Sepas-Moghaddam *et al*. [14] introduced a three-branch feature extraction method to utilize the multi-modal complementary discriminative information, whereby the RGB central view, depth map, and disparity map were extracted from the light-field image using three VGG branches. The extracted high-level features were then fused and classified by SVM. Uppal *et al*. [15] proposed an attention mechanism that used the depth map to guide the attention mechanism to selectively focus on important parts of the input RGB image.

Existing face recognition methods on RGB-D images usually use two branches to extract features from RGB and depth images for attention. Three-branch approach with an additional one-channel disparity map has been studied in [14], but they only utilized feature-level fusion without attention. Since the availability of light-field raw datasets is very limited, they only evaluated on a small dataset with 100 subjects, and each subject contains merely 20 images. Its ablation study showed that the disparity map is less effective when it was used alone with RGB. However, when it was used together with depth in addition to RGB, the recognition accuracy increased. This has motivated us to explore the effective source of single-channel images in conjunction with the depth map. In [16], LBP calculated from RGB is used to guide attention for facial expression recognition. It showed that two-dimensional textures described using local spatial patterns and gray-scale contrast can provide better attention. LBP is robust towards monotonic gray-scale changes such as illumination variations. Starting from any arbitrary point and direction (*i.e*. clockwise or counterclockwise), the LBP of a central pixel (x_c, y_c) can be efficiently calculated from 8 points in its 3×3 neighborhood as follows:

$$\text{LBP}\left(x_c, y_c\right) = \sum_{p=0}^{7} S\left(i_p - i_c\right) 2^p, \tag{4}$$

where S is the sign function.

Compared with other features, LBP is more effective in guiding RGB-D face recognition based on the following three criteria: 1) LBP and its variations have favorable characteristics such as gray-scale and rotation invariance for resolving illumination and rotation angle problems, which are critical in face recognition scenario; 2) As a texture extractor, LBP can reflect subtle discriminative changes, and is easy to compute. These make it suitable for resource-constrained applications that require texture features; 3) LBP has shown to be effective in guiding the attention module, especially in facial expression recognition [16].

Facial expression recognition and face recognition share some commonalities with different goals. The former recognizes the expression regardless of the identity, whereas the latter attempts to correctly predict the identity under facial variations of the same person. In cognition of their subtle differences, we propose a CNN-based three-branch neural network for RGB-D face recognition for the first time. This combination of LBP features and depth map for texture- and depth-guided attention mechanism has not been conceived before as a means to greatly enhance the robustness of the face recognition learning model on covariate changes.

III. Proposed Architecture and Optimization

A. Network Architecture

The proposed network is shown in Fig. 1. It consists of three modules, *i.e*., feature extraction module with convolutional filters, attention fusion module for embedding extracted features from three separate branches, and attention refinement module that combines attention map and face features for the final identification.

Convolutional feature extraction module uses three CNN branches to extract features from the input images. We introduce an LBP branch into the neural network to jointly enhance attention with the depth map. The LBP images are calculated from RGB using Eqn. (4). The extracted feature maps in (M, M, Φ) are reshaped to $(M * M, \Phi)$ and passed to the attention fusion module for feature-level fusion and feature pooling.

Attention fusion module is the core of our architecture which merges the output from the three feature extractors to guide the attention map generation. Compared with two-branch models [14], [16], our proposed model has an additional LBP branch for extra discriminative information. As

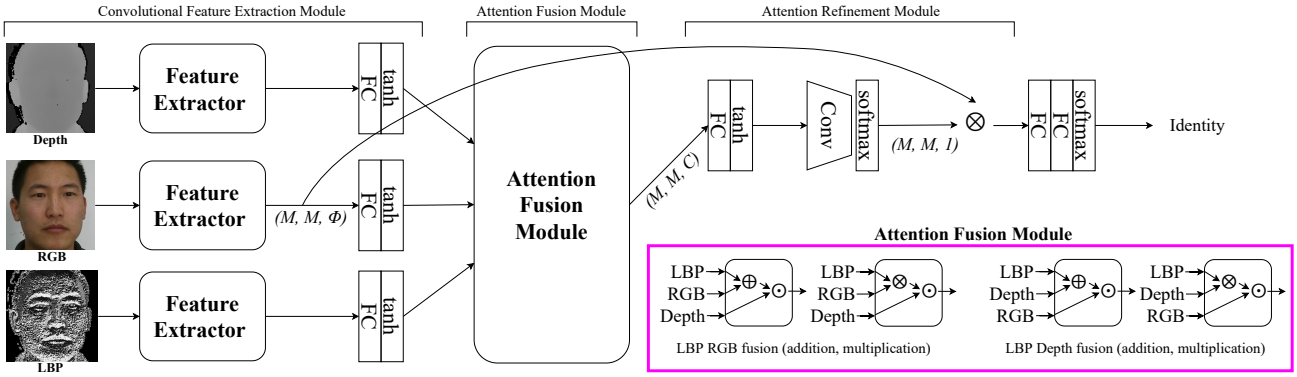


Fig. 1. Architecture of the proposed model

highlighted in Fig. 1, a two-modal joint-representation F_{fusion} for two selected branches is first obtained using Eqn. (1) or Eqn. (2). It is then pooled with the remaining branches to find their interactions and produce F_{fusion} using dot product. The information from LBP can be either utilized to highlight the fine changes on RGB or use in conjunction with depth to maximize interactions. In the attention fusion module, there are two approaches: 1) fuse with depth and pool with RGB; 2) fuse with RGB and pool with depth. Since LBP is computed from RGB, the similarity between these two modalities may result in redundant information in fusion. Therefore, a better option is to first fuse LBP and depth as they are independent of each other. Interaction between the fused features and the remaining modality is then found using feature pooling.

Attention refinement module is adapted from [15]. Non-linearity is added from the FC layer with a tanh activation. An attention map in $(M, M, 1)$ is then generated using 1×1 convolution. The fused and pooled feature map is resized to (M, M, C) and passed to the attention refinement module for refining the attention on RGB features. Categorical cross-entropy loss is used for the minimization of all three branches of the network.

As opposed to two-branch models that only perform feature-level fusion of RGB and depth features, the attention fusion module of our proposed three-branch model explores different combinations of feature-level fusion and pooling mechanisms by utilizing information from the extracted LBP in conjunction with RGB and depth.

B. Model Optimization for Edge Deployment

In deep learning, accuracy is the primary performance criterion. However, accuracy improvement is a challenging task as it is often accompanied by increasing model complexity and computational cost, which limits its implementation feasibility on resource-constrained edge devices. In multi-modal face recognition, the facial identity is mainly derived from RGB data, while auxiliary modalities like depth and LBP provide additional information to improve accuracy and robustness under challenging environments. Compared with two-branch methods, the introduction of the LBP branch increases the number of feature extractors by one, resulting in higher model complexity.

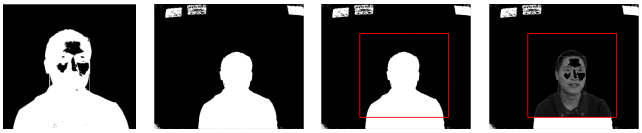
The convolutional feature extraction module is the most compute-intensive component of the network. In our proposed method, depth and LBP are only used for attention refinement. Thus, by reducing the complexity of these two branches, the model efficiency can be improved with negligible degradation on the already attained improved accuracy. Provided that the output dimensions of the three feature extractor branches are kept the same, the rest of the existing network structure can be reused. The effect of reducing the complexity of the two auxiliary branches and the RGB branch is evaluated in Section IV. We first use VGG as the feature extractor for all three branches to achieve the best baseline accuracy. We then conduct rigorous experiments to empirically replace both the auxiliary branches and RGB branch with MobileNet v3 as feature extractors for lightweight and faster inference on edge device deployment.

IV. Experiment Results

A. RGB-D Datasets

We evaluated the proposed model on two popular RGB-D datasets, Lock3DFace and IIIT-D. Lock3DFace is a large-scale dataset consisting of 5671 RGB-D face video clips of 509 individuals with diverse changes. This dataset was recorded in two sessions. Each session contains four categories: neutral, pose variation, facial expression, and occlusion. The number of images in each category from one session is around 40000, and the total number of images for Lock3DFace dataset is around 320000. Only neutral images from the first session are used as the training set. The remaining three variations of the first session form three test sets for pose, occlusion, and facial expression. All the images from the second session, including both neutral and variations, form the fourth test set. IIIT-D dataset contains RGB-D images from 106 subjects in two sessions. The images were captured under normal illumination with variations in pose, expression, and eyeglasses. The dataset was preprocessed to keep only the face regions and split into 5 folds for a predefined 5-fold cross-validation test protocol. For each set, there are around 4200 images for training and 1400 images for validation.

It is worth noting that the RGB and depth images in Lock3DFace dataset were not aligned with the camera parameters during acquisition. Similar issues can be found in some other datasets [16] acquired using also Microsoft Kinect.



(a) RGB (b) Depth (c) Matched ROI (d) Overlapped

Fig. 2. Thresholding and template-matching results

TABLE I
ACCURACY COMPARISON OF STATE-OF-THE-ART FACE RECOGNITION MODELS AND PROPOSED METHOD ON THE LOCK3DFACE DATASET

Model	Input	Accuracy			
		Pose	Expression	Occlusion	Session2
ResNet50 [15]	RGB	58.4%	96.3%	74.7%	75.5%
Inception-v2 [17]	RGB-D	54.6%	97.3%	69.6%	66.1%
CNN-MFCC-SAV [18]	Mapping	70.4%	98.2%	78.1%	65.3%
Depth Attention [15]	RGB-D	64.9%	97.8%	84.8%	78.0%
LBP-depth Attention	RGB-D	66.4%	98.1%	87.6%	79.7%

As no preprocessing guidelines are given, we propose a template-matching-based alignment method to automate data preprocessing. First, the face in the RGB image is detected, and the region of interest (ROI) is widened to include more parts (*i.e.* neck, shoulder) of the body to improve matching accuracy. Next, the depth image and the cropped RGB image are thresholded to obtain the masks. By searching for the scaled RGB mask in the depth mask, the corresponding ROI in the depth mask can be determined. Finally, the obtained ROIs in the RGB and depth masks are shrunk to the original size of the detected face and used to crop out the RGB and depth face images. This alignment process is shown in Fig. 2.

B. Evaluation Results

The performances of recent state-of-the-art models and our proposed model on the Lock3DFace dataset are shown in Table I. Compared with RGB input model [10], our network has higher overall recognition accuracy, especially on pose and occlusion scenarios. Similarly, our model exceeds the accuracy of the reconstructed input model [18] by $\sim$10% in occlusion and session2 tests. It also outperforms Inception-v2 + fusion [17] method on RGB-D images by 11.8%, 0.8%, 18% and 13.6% in pose, expression, occlusion and session2, respectively. We replicated the depth attention method [15] based on our template alignment process. In comparison, our model has better accuracy on all the test sessions. Compared with Lock3DFace dataset, IIIT-D dataset is smaller in scale and contains less challenging variations. The accuracy results on IIIT-D dataset are given in Table II. Our model obtains the best accuracy compared with other methods including SRC, SVM classifiers, two-level attention, and depth-guided attention.

To examine and compare the effectiveness of our proposed attention model against depth-guided attention [15], GradCAM is used for visualizing the attention maps on Lock3DFace test set images as shown in Fig. 3. The left column of each subfigure represents the attention maps from [15] and the right column shows the attention map obtained by our proposed network. Compared with using only depth information to guide the attention, the GradCAM attention heatmap images show that our LBP-depth-guided attention has a better ability to guide the model to extract discriminative features on important

TABLE II
ACCURACY COMPARISON OF RECENT STATE-OF-THE-ART FACE RECOGNITION MODELS AND THE PROPOSED METHOD ON THE IIIT-D DATASET

Model	Classifier	Input	Accuracy
LBP + Haar + Gabor [19]	SRC	RGB-D	91.3%
Covariance Matrix [20]	SVM	RGB-D	96.4%
Two-level Attention [21]	FC	RGB-D	98.2%
Depth Attention [15]	FC	RGB-D	99.1%
LBP-depth Attention	FC	RGB-D	100%

(a) Facial expression (b) Pose variation (c) Occlusion

Fig. 3. GradCAM attention map on the test set of Lock3DFace dataset. Depth-guided attention results are on the left, and LBP-depth-guided attention results are on the right of each pair of images.

regions such as the eyes, nose, mouth, and eyebrow. In Fig. 3, for all three challenging cases, our model has more accurate coverage of the discriminative regions. The results indicate that the fine details supplied by LBP can guide the model to learn more meaningful feature representatives from the neutral faces with better generalization abilities on unseen data.

C. Deployment on IoT devices

Neural network inference is increasingly pushed to the edge to bring analytics closer to the data source. In this experiment, we deployed our trained model on a Jetson Nano 4GB device. The Tegra X1 chipset on the board allows model execution using CUDA, which outperforms CPU inference in terms of speed and power efficiency.

To run inference on edge platforms with limited computing and power budget, many lightweight model architectures have been proposed, such as SqueezeNet [22] and MobileNet [23]. In this experiment, we evaluate different feature extractor combinations on Lock3DFace dataset using VGG and MobileNet v3. The model accuracy and inference speeds of different feature extractor combinations and fusion strategies are summarized in Table III. VGG-based models have better performance on occlusion and session2 which have many images with rich variations. For facial expression, since the variations relative to the neutral face images are small, the performance of MobileNet-based models is comparable, if not better than VGG-based models. Interestingly, the model with VGG for the RGB branch and MobileNet v3 Small for depth and LBP branches achieved the best performance on the pose, and the accuracy degradations on the other three groups are also minimal ($<$1%).

Our experiment shows that the complexity of the RGB feature extractor has a huge impact on recognition accuracy, while less complex feature extractors are still capable of guiding attention and providing satisfactory results. When depth and

TABLE III

ACCURACY AND FRAME RATE COMPARISON OF DIFFERENT FEATURE EXTRACTOR AND ATTENTION FUSION STRATEGY COMBINATIONS (LOCK3DFACE). OC DENOTES OVERCLOCKING (CPU@1.85GHZ, GPU@0.95GHZ) AND D DENOTES DEFAULT FREQUENCY (CPU@1.479GHZ, GPU@0.9216GHZ).

RGB branch	Depth branch	LBP branch	Attention fusion strategy	FPS (Default/OC)	Accuracy			
					Pose	Expression	Occlusion	Session2
VGG			Depth + LBP	1.61 / 1.66	66.62%	97.80%	87.03%	79.37%
VGG			Depth × LBP		66.43%	98.12%	87.62%	79.68%
VGG			RGB + LBP		66.22%	98.21%	84.19%	78.55%
VGG			RGB × LBP		66.09%	97.82%	85.26%	77.95%
MobileNet v3 Large			Depth + LBP	8.0 / 8.7	47.92%	98.70%	86.30%	75.18%
MobileNet v3 Large			RGB + LBP		45.81%	98.68%	84.93%	74.46%
VGG	MobileNet v3 Large		Depth + LBP	3.4 / 3.5	65.82%	98.26%	85.96%	78.90%
VGG	MobileNet v3 Large		Depth × LBP		65.85%	97.56%	85.61%	78.90%
VGG	MobileNet v3 Small		Depth + LBP	3.8 / 4.0	69.14%	97.96%	84.32%	79.04%
VGG	MobileNet v3 Small		Depth × LBP		68.65%	97.95%	86.52%	78.99%

LBP branches are replaced with MobileNet v3, either Large or Small, the accuracy drop is very small compared with using all VGG branches. The accuracy can possibly be increased in some cases. Therefore, when deploying multi-modality models on IoT devices, reducing the complexity of less discriminative modalities is a viable solution for efficiency with minimal accuracy loss.

V. CONCLUSION

This paper presents a novel LBP-depth-guided attention model for RGB-D face recognition. Both quantitative and qualitative experiment results indicate that LBP-depth-guided attention has a better focus on the important discriminative regions under challenging imaging conditions. By fusing LBP with depth and pooling the fusion result with RGB, we successfully improved the recognition accuracy under various scenarios including occlusion and pose variations. In addition, we also successfully deployed our proposed model on Jetson Nano with CUDA acceleration. Several feature extractor combinations for model complexity reduction were also explored on the same platform to demonstrate the adaptivity of the proposed LBP-depth-guided attention method.

ACKNOWLEDGMENT

This research is supported by the Ministry of Education, Singapore, under its AcRF Tier 2 Award No. MOE-T2EP50220-0003.

REFERENCES

[1] A. Kumar, P. S. Kumar, and R. Agarwal, "A face recognition method in the iot for security appliances in smart homes, offices and cities," in *Proc. 3rd Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Erode, India, Mar. 2019, pp. 964–968.

[2] J. Kittler, A. Hilton, M. Hamouz, and J. Illingworth, "3d assisted face recognition: A survey of 3d imaging, modelling and recognition approachest," in *Proc. IEEE Comput. Vision Pattern Recognit. (CVPR)*, San Diego, CA, USA, Sep. 2005, pp. 114–114.

[3] Z. Hu, P. Gui, Z. Feng, Q. Zhao, K. Fu, F. Liu, and Z. Liu, "Boosting depth-based face recognition from a quality perspective," *Sensors*, vol. 19, no. 19, p. 4124, Sep. 2019.

[4] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3dface: A large-scale database of low-cost kinect 3d faces," in *Proc. IEEE Int. Conf. Biometrics (ICB)*, Halmstad,Sweden, Jun. 2016, pp. 1–8.

[5] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On rgb-d face recognition using kinect," in *Proc. IEEE 6th Int. Conf. Biometrics: Theory, Appli. Syst. (BTAS)*, Arlington, VA, USA, Sep. 29 - Oct. 2 2013, pp. 1–6.

[6] G. Goswami, M. Vatsa, and R. Singh, "Rgb-d face recognition with texture and attribute features," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1629–1640, Oct. 2014.

[7] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[8] B. Y. Li, A. S. Mian, W. Liu, and A. Krishna, "Using kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Workshop Appl. Comput. Vision (WACV)*, Clearwater Beach, FL, USA, Jan. 2013, pp. 186–192.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," May 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogni. (CVPR)*, Las Vegas, Nevada, USA, Jun. 2016, pp. 770–778.

[11] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1449–1457.

[12] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. European Conf. Comput. Vision (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 451–466.

[13] A. Chowdhury, S. Ghosh, R. Singh, and M. Vatsa, "Rgb-d face recognition via learning-based reconstruction," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Niagara Falls, NY, USA, Sep. 2016, pp. 1–7.

[14] A. Sepas-Moghaddam, P. L. Correia, K. Nasrollahi, T. B. Moeslund, and F. Pereira, "Light field based face recognition via a fused deep representation," in *Proc. IEEE 28th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Aalborg, Denmark, Sep. 2018, pp. 1–6.

[15] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Depth as attention for face representation learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2461–2476, Jan. 2021.

[16] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "Attention mechanism-based cnn for facial expression recognition," *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020.

[17] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2d face recognition via discriminative face depth estimation," in *Proc. IEEE IEEE Int. Conf. Biometrics (ICB)*, QLD, Australia, Feb. 2018, pp. 140–147.

[18] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang, "Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces," in *Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogni. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5773–5782.

[19] B. Y. Li, A. S. Mian, W. Liu, and A. Krishna, "Face recognition based on kinect," *Patt. Anal. Appl.*, vol. 19, no. 4, pp. 977–987, Nov. 2016.

[20] M. Hayat, M. Bennamoun, and A. A. El-Sallam, "An rgb–d based image set classification for robust face recognition from kinect data," *Neurocomputing*, vol. 171, pp. 889–900, Jan. 2016.

[21] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Two-level attention-based fusion learning for rgb-d face recognition," in *Proc. IEEE 25th Int. Conf. Patt. Recogni. (ICPR)*, Milan, Italy, Jan. 2021, pp. 10 120–10 127.

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[23] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Seoul, Korea, Oct. 27th - Nov. 2rd 2019, pp. 1314–1324.