Median Pixel Difference Convolutional Network for Robust Face Recognition

Jiehua Zhang[1]
jiehua.zhang@oulu.fi

Zhuo Su[1]
zhuo.su@oulu.fi

Li Liu*,[2,1]
li.liu@oulu.fi

[1] Center for Machine Vision and Signal Analysis
University of Oulu
Finland

[2] College of System Engineering
National University of Defense Technology
China

Abstract

Face recognition is one of the most active tasks in computer vision and has been widely used in the real world. With great advances made in convolutional neural networks (CNN), lots of face recognition algorithms have achieved high accuracy on various face datasets. However, existing face recognition algorithms based on CNNs are vulnerable to noise. Noise corrupted image patterns could lead to false activations, significantly decreasing face recognition accuracy in noisy situations. To equip CNNs with built-in robustness to noise of different levels, we proposed a Median Pixel Difference Convolutional Network (MeDiNet) by replacing some traditional convolutional layers with the proposed novel Median Pixel Difference Convolutional Layer (MeDiConv) layer. The proposed MeDiNet integrates the idea of traditional multiscale median filtering with deep CNNs. The MeDiNet is tested on the four face datasets (LFW, CA-LFW, CP-LFW, and YTF) with versatile settings on blur kernels, noise intensities, scales, and JPEG quality factors. Extensive experiments show that our MeDiNet can effectively remove noisy pixels in the feature map and suppress the negative impact of noise, leading to achieving limited accuracy loss under these practical noises compared with the standard CNN under clean conditions.

1 Introduction

Face Recognition is one of the most important research fields in computer vision and pattern recognition. Recent advances in deep learning, coupled with abundant face data, have led to excellent progress in face recognition algorithms [8, 24, 27, 30, 31, 32, 37]. Due to these achievements, face recognition technology is widely utilized in the real world, such as human-computer interaction[20], video surveillance[5], and identification [14, 34]. The accuracy is a crucial metric to evaluate the effectiveness of recognition models. However, the existing noise produced by the sensor of a scanner, cameras, image compression, *etc.*, results in a significant decrease in the accuracy of models. Enhancing the noise-robustness is meaningful to the practical applications of face recognition.

* indicates the corresponding author.

Recently, many studies have been carried out to deal with the challenges of face recognition, such as the variations in expressions, lighting conditions, and poses [4, 16, 19], with few efforts taking noise into account. The existing noise-robust face recognition methods mainly consider the noisy labels in the face dataset [9, 10, 33, 36, 44]. This kind of method is hard to adapt to the diversity and complexity of noise in the real world. Our purpose is to improve the robustness of face detectors to noise in the real world.
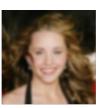
In this paper, we propose an effective approach to enhance the noise-robustness of CNNs with limited decreasing accuracy compared with standard CNNs. The existing CNNs exhibit sensitivity to noise, particularly salt-and-pepper noise. Inspired by [21], we proposed a robust Median Pixel Difference Convolutional Network (MeDiNet), which can effectively suppress noise and improve the performance under noisy conditions by median pixel difference convolution (MeDiConv). Specifically, MRELBP replaces the individual pixel intensities in the feature maps with median representations in the corresponding local regions and reports excellent results in the texture classification task under various noises, showing that combining local median representation is powerful to extract features. The proposed MeDiConv inhabits noise-robust property of MRELBP.

We summarize our contributions as follows: (i) We proposed a noise-robust network MeDiNet integrating a novel convolutional operation named as MeDiConv to utilize median representation to suppress noise in images effectively. (ii) We adopt the image degradation model with a broad range of noises to approximate the noisy model in the real world (blur kernel, noise intensity, scaling, JPEG quality), which can evaluate the performance of our method in practical application. (iii) The model can be robust to a variety of noises without generating additional noisy images for training the model.

In the experiment, MeDiNet is tested on four face datasets with versatile settings. For approximating the complex noise model in the real world, we also add two complex noises to evaluate our method except salt-and-pepper noise and additive white Gaussian noise (AWGN): Heteroscedastic Gaussian (HG) [26] and multivariate Gaussian (MG) noise [43]). The noisy image examples are shown in Figure 1. Our experimental results show that the MeDiConv can effectively suppress the negative impact of the noise on the images. We demonstrate that MeDiNet reports superior performance to standard CNN architectures on four face datasets under noisy conditions (LFW [12], CA-LFW [46], CP-LFW [45], Youtube Face [38]) while the loss of accuracy on the clean images is limited.

(a) Clean Image   (b) Gaussian Blur   (c) Gaussian Noise   (d) JPEG Compression

Figure 1: The image example with versatile settings. (b) Gaussian blur, σ=3, kernel size=13; (c) Gaussian noise, σ=20; (d) JPEG Compressor, JPEG quality=10.

2 Related work

2.1 Face recognition in noisy conditions

With the development of deep learning, a large number of face recognition algorithms have been proposed to address problems like lighting, age, pose [4, 45, 46]. Noise in human face images can significantly decrease the performance of face recognition systems. Noise drops the valuable information in the image and influences the ability of some algorithms to recognize a correct object. To suppress the impact of noise, Ding *et al.* [7] proposed the NR-Network for face recognition under noise, which utilized a multi-input structure to extract a multi-scale and more discriminative feature from the input image. Wu *et al.* [39] proposed a denoising network D-BSN that is designed as a two-stage training framework with self-supervised learning and knowledge distillation to learn the denoising from unpaired clean and noisy images. Anwar *et al.* [2] firstly incorporate an attention mechanism to exploit channel dependencies in the denoising task. Li *et al.* [18] proposed GFRNet consisting of warping and a reconstruction network. The GFRNet combined the degraded and high-quality images to enhance the performance of face restoration. The existing studies for suppressing noise focus on the restoration tasks [1, 2, 6, 18, 39, 41]. These methods depend on training with noisy and clean images. We aim at designing a general framework without extra noisy image data to suppress noise in the face recognition task directly.

2.2 LBP and convolution

LBP [25] was proposed for texture analysis and has been successfully applied to many tasks [22], such as image retrieval [35], object detection [23] and face image analysis [11, 28]. LBP has emerged as one of the most prominent texture descriptors. Recently, some studies tried to combine LBP with standard CNNs [29, 42]. The LBCNN was motivated by LBP and utilized a set of fixed sparse pre-defined binary convolutional filters to achieve computational savings [13]. Inspired by LBCNN, the LBVCNN added the LBV layer to reduce trainable parameters in facial expression task [15]. Li *et al.* [17] combined the LBP and attention mechanism to extract useful features, which achieved higher performance in face recognition. These studies demonstrated that combining LBP with standard CNNs is a practical and flexible approach in various tasks. The MRELBP is based on combining a median filter with LBP operation in a multiscale fashion, which can effectively extract noise-resistant patterns. The development of MeDiNet shares similar motivations but has slightly different implementation due to the different processing procedures of LBP and CNN.

3 MeDiNet

The MeDiNet is built on the Sphere20 [24]. The initial convolutional layers of Sphere20 are replaced with MeDiConv for suppressing the negative impact of a broad range of noises in the face images. The operation of MeDiConv is introduced as follows.

3.1 Preliminary of MRELBP

The traditional LBP calculates the differences between the pixel value of the central point and those of its neighbors. The MRELBP [21] set the pixel intensity as median representation
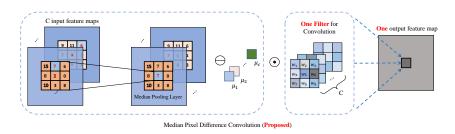
Figure 2: The convolutional operation of Median Pixel Difference Convolution. μ_i denotes a global predefined value of each channel.

in a local patch, which is more robust to gray scale variations, rotation changes, and noise. We choose the MRELBP descriptor under center pixel representation, a simple and effective way to handle noise. This descriptor can be expressed as:

$$MRELBP\_CI(x_c) = s(\phi(X_{c,w}) - \mu_w) \tag{1}$$

where s is the sign function, $\phi(X_{c,w})$ denotes the median operation in $X_{c,w}$ (the local patch of size $w \times w$ centered at center pixel x_c), μ denotes the mean value of $\phi(X_{i,w})$ over the whole image X_i. Inspired by this work, we use a set of learnable weight parameters to replace sign function for combining the standard CNN with median representation. The MRELBP is based on combining a median filter with traditional LBP operation, which can effectively extract powerful features under various noisy conditions. Our MeDiConv also utilize median representation in convolutional operation. This processing can make convolutional layer suppress noisy pixels in forward propagation.

3.2   MeDiConv

The calculation process of standard CNN can be described as: the convolutional kernel $W^{c_{out} \times c_{in} \times k \times k}$ slides on the feature map $X^{c_{in} \times H \times W}$ to calculate the output. The formulation can be expressed as following (c_{out} is set to 1 without loss of generality).

$$X_{out}(u,v) = \sum_{i=0}^{c_{in}-1} \sum_{(p,q)\in\mathcal{N}} X_i(u+p,v+q) \cdot W_i(p,q) \tag{2}$$

where (u,v) denotes the coordinate of the pixel in feature map, W_i denotes the convolutional kernel, and $\mathcal{N}$ denotes the $k \times k$ neighborhood area centered at (u,v). The standard CNNs are sensitive to noise (especially salt-and-pepper noise) since the noise in the local region can confuse the subsequent feature representation process and lead to the detection of false features with no physical correlation.

To suppress noise in regions of feature maps, MeDiConv utilizes local median pixel difference value to replace the original value of $x_{u,v}$ in the convolutional operation. As shown in Figure 2, the value of each pixel x_i in the input feature patch is replaced by the median value of the 3x3 neighborhood centered on x_i. The procedure can be decomposed as the following stages: 1) for each location (u,v) with its neighborhood region $\{(u+p,v+q)|(p,q)\in\mathcal{N}\}$, calculating the median representation for each pixel located in the neighborhood region; 2)

replacing the pixel intensity $X_i(u+p, v+q)$ in Eq. 2 with the difference between the calculated median representation of $X_i(u+p, v+q)$ and a global predefined value μ; 3) combining the output in stage 2 with the weights to generate the output feature map. Therefore, MeDiConv can be expressed as:

$$\tilde{X}_{out}(u,v) = \sum_{i=0}^{c_{in}-1} \sum_{(p,q)\in\mathcal{N}} (\phi(X_i(u+p, v+q)) - \mu) \cdot W_i(p,q) \tag{3}$$

$$\phi(X_i(u,v)) = Median(\{X_i(u+p, v+q)|(p,q)\in\mathcal{N}\}) \tag{4}$$

where $\tilde{X}_{out}$ is the output feature map, $\phi(X_i(u,v))$ denotes the median representation of the $k \times k$ region centered at $X_i(u,v)$, and μ represents the average value of $\phi(X_i(u,v))$ over the whole feature map. The aim of subtracting μ is to eliminate the bias of pixel intensities in the feature map.

Compared with standard convolution, the MeDiConv is a nonlinear smoothing operation, which can effectively remove outliers with limited impact on the ability of feature extraction [5]. Noise corrupted image patterns could lead to false activations, leading to a significant decrease in accuracy. For MeDiConv, the effect of applying MeDiConv at multiple layers can be considered as applying multiple median filters of different kernel sizes on the original image as each MeDiConv layer has a different receptive field. That is also why the proposed MeDiNet can deal with noise of different levels. Considering the trade-off between accuracy and noise robustness, MeDiNet only set MeDiConv in the first few layers. When the standard CNN in the deep layer of the network is replaced by MeDiConv, the response of the feature map will be degraded. Setting MeDiConv in the beginning layers is sufficient to eliminate the impact of noise.

4 Experiments

4.1 Dataset

We adopt the CASIA-WebFace [40] as a training dataset. The LFW [12], CA-LFW [46], CP-LFW [45], Youtube Face [38] are used for evaluation. The details of datasets are as following:

CASIA-WebFace. The CASIA-WebFace dataset is collected from IMDb website containing 10,575 subjects and 494,141 images. The collection and annotation of CASIA-WebFace use a semi-automatical way. It also utilizes the name bag and face similarity for efficient labeling.

LFW. Labeled Faces in the Wild (LFW) aims to study face recognition in an unconstrained environment, containing more than 13,000 face images from the website. Each face in the LFW is marked with a person's name, and about 1680 people contain more than two faces. Although there are some mislabeled matched pairs, LFW is considered completely clean.

CA-LFW and CP-LFW. CA-LFW and CP-LFW are renovations of LFW dataset. These datasets focus on the cross-age and cross-pose face challenges in face recognition. The images are collected from 3,000 positive face pairs with age gap and pose gap, respectively. Compared with the LFW, both datasets are more challenging with lower test accuracy.

YouTube Faces. The YouTube Faces Dataset (YTF) is a facial video database used to study unconstrained face recognition in videos. The dataset contains 3,425 videos from

1,595 different people. All the videos are collected from YouTube. Each subject provides an average of 2.15 videos. The shortest clip length is 48 frames, the longest clip length is 6070, and the average video clip length is 181.3.

4.2   Degradation model

In the practical application, the quality of images can be affected by noise, scaling, compression, and their combination. For verifying the robustness of MeDiNet, a degradation model is required to generate realistic noisy images [18], which can be defined as:

$$I^d = ((I \otimes k)_{\downarrow_S} + n)_{JPEG_C} \tag{5}$$

where $\otimes$ means the convolutional operation. k denotes the blur kernel. $\downarrow_S$ denotes the downsampling operation with the scale factor S. n denotes the noise like AWGN. $(\cdot)_{JPEG_C}$ denotes the JPEG compression with the quality factor c. The degradation $I^d = ((I \otimes k)_{\downarrow_S} + n)$ is caused by remote acquisition and $(\cdot)_{JPEG_C}$ expresses the effect of JPEG compression.

The parameter settings of our degradation model is shown as below:

Blur kernel. In our paper, we utilize Gaussian blur kernel and Motion blur kernel as the Simulation of defocusing effect. The variance ρ of Gaussian blur is sampled from the set $\{1 : 1 : 3\}$. The degree $\mathcal{D}$ of Motion blur is sampled from the set $\{10 : 10 : 40\}$. 0 means no blur operation.

Downsampler. The scale factor $\mathcal{S}$ is sampled from the set $\{2, 4, 6, 8\}$. It should be noted that the image will be upsampled to its original size after downsampling.

Noise. Besides common noise (like AWGN and salt-and-pepper noise), we select more complex noise to evaluate our model, including HG $n_i \sim \mathcal{N}(0, \alpha^2 x_i + \delta^2)$ and MG noise $n \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \mathcal{L}^2 \cdot U \Lambda U^T$. U is a random unitary matrix, Λ is a diagonal matrix of three random values in the range $(0,1)$. The sampling ranges of noisy parameters are following: 1) AWGN: σ in $\{15, 25, 50\}$; 2) Salt-and-pepper noise: ρ in $\{5\%, 10\%, 15\%, 25\%\}$; 3) HG noise: α in $\{20 : 10 : 40\}$, σ in $\{10, 20\}$; 4) MG noise: $\mathcal{L} = 75$.

JPEG compression. When storing images, the quality would be compressed. We add the JPEG Compression with the quality factor $\mathcal{C}$ on the degraded images. The $\mathcal{C}$ is sampled from $\{10 : 10 : 40\}$. 0 means the image is losslessly compressed.

4.3   Experimental settings

Our model is trained on the CASIA-WebFace dataset. We crop the training image with size 112x96 to feed the network. The model is trained in 20 epochs with SGD optimizer. The batchsize is set as 256. The learning rate is initialized as 1e-2 and decayed by factor 0.1 after 10, 15, 18 epoch. We add the degradation model on test images for evaluation and conduct a separate experiment to analyze the influence of independent noise.

4.4   Experiments on the images under single noise

Our main objective of this work is to enhance the inherent noise robustness of CNNs without significant performance loss. Therefore, Our method focuses on applying median representation to the CNN. We set the Sphere20 as baseline [24] and compare the performance of different MeDiConv layers. The MeDiNet is built on the Sphere20. The first to fifth layers of Sphere20 are replaced with the MeDiConv (For example, the MeDi3 denotes the first three

layers are MeDiConv). The training stage only used clean images. The noisy images were used in the test stage. Various types of noise are utilized to evaluate our training model on the LFW and YTF datasets (See Figure 3 and Figure 4). We separately analyzed the robustness of MeDiNet to each noise type. *The detailed information of results can be observed in the supplemental material.*

Evaluation on blurring. In practical applications, the blurred images are generated from defocus and pre-processing stage. We evaluate MeDiNet on the gaussian and motion blur kernel. The blurring leads to the degradation of edges and structures in the image. The results illustrate that our method is robust to image blurring. However, the performance drops slightly with the number of MeDiConv increasing to five layers.

Evaluation on AWGN and Salt-and-Pepper noise. These noises are generated from remote sensing and can be regarded as outliers in the images. The standard CNNs are sensitive to noise. Noise pixels can significantly reduce the accuracy of face recognition, especially salt-and-pepper noise. When setting the proportion of salt-and-pepper noise to 20%, the accuracy of model drops by more than 30%. MeDiNet can effectively remove the outliers in the feature maps by median representation. The test results in Figure 3 and Figure 4 show that our method can enhance the robustness to a broad range of noises.

Evaluation on complex noise. However, CNN usually generates poorly due to complex noise even if it can adapt to AWGN and salt-pepper noise [26]. Thus, we also present the results on complex noise models: HG and MG noise (See Figure 3 and Figure 4). We can observe that the improvement is still significant, which can prove the robustness of proposed MeDiNet.

Evaluation on Scaling and Compression. Standard CNNs achieve slight accuracy loss under compression. In the procedure of JPEG compression, part of high-frequency information is removed, which has a negative impact on the ability of CNN to extract features. For scaling, results demonstrate that the accuracy drops significantly when enlarging scaling factor S because of sampling and interpolation. The detailed information of images might be removed during the scaling process. The median representation can effectively enhance the expression ability of features to a certain extent.
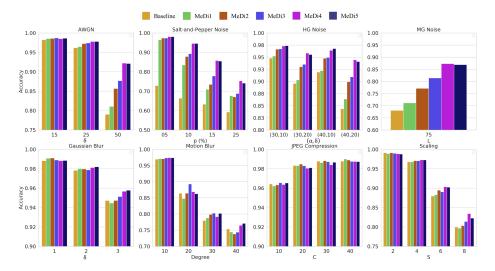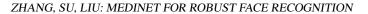

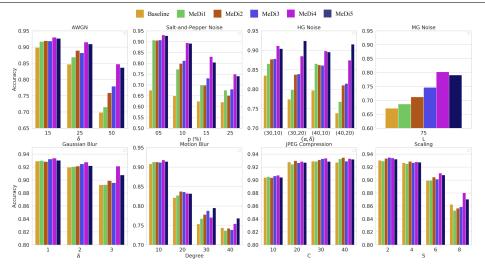Figure 3: The evaluation on the LFW dataset under single noise.

Figure 4: The evaluation on the YTF dataset under single noise.

Through the analysis of each experimental result under the single noise, we can observe that the MeDiNet inherits the characteristics of MRELBP, and demonstrates striking robustness to various image noises, including blurring, AWGN, Salt-and-Pepper, HG, MG, Compression and Scaling. In the Sections 4.5, the combination of multiple noises is adopted to test our method, which approximates noise interference in the real world. We visual-
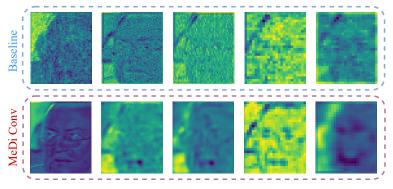

Figure 5: The visualization of feature map. The first row are outputs of the first to fifth layer in the baseline; the second row are outputs of the last MeDiConv layer in the MeDi1-5 (*e.g.,* the second figure denote the feature map of last MeDiConv layer in MeDi2).

ized feature maps of middle layers in the model to demonstrate the ability of MeDiConv to suppress noise (See Figure 5). The input image was added with the AWGN with $\sigma = 50$. We can observe that the output by MeDiConv is smooth and contains fewer noisy pixels. The structure and edges of face are identifiable, leading to improve the accuracy of face recognition.

4.5 Experiments on the degraded images

In this section, the degradation model introduced in Sections 4.2 is adopted to approximate the real-world noise. We utilize two blurring methods and four types of noise to combine.

The parameters of the degradation model are randomly generated within their respective parameter ranges in Sections 4.2 (*The value of parameters can be observed in the supplemental material*). In this section, we present the evaluation of four combining noises:

- Gaussian_blur+AWGN+Scaling+Compression (GB+AWGN);
- Gaussian_blur+Salt-and-Pepper+Scaling+Compression (GB+SP);
- Gaussian_blur+HG+Scaling+Compression (GB+HG);
- Gaussian_blur+MG+Scaling+Compression (GB+MG).

The *clean* denotes no noise in the images. Under real-world noisy conditions, the face recognition accuracy drops more significantly than it under single noise. Tab.1 and Tab.2 show the evaluation of four datasets under real-world noisy conditions. *The results of other combinations can be observed in the supplemental material.*

Table 1: The evaluation on the LFW and YTF dataset

	Clean		GB+AWGN		GB+SP		GB+HG		GB+MG	
	LFW	YTF	LFW	YTF	LFW	YTF	LFW	YTF	LFW	YTF
Sphere20	**0.9922**	0.9330	0.7682	0.7636	0.6762	0.6618	0.7482	0.7440	0.6573	0.6628
MeDi1	0.9908	0.9300	0.7938	0.7946	0.7407	0.7472	0.7790	0.7500	0.6723	0.6790
MeDi2	0.9910	0.9322	0.8390	0.8256	0.7797	0.7848	0.8192	0.7984	0.6755	0.6754
MeDi3	0.9898	0.9328	0.8402	0.8254	0.7908	0.7888	0.8208	0.8058	**0.7265**	0.7178
MeDi4	0.9882	**0.9350**	**0.9083**	**0.8864**	**0.8820**	**0.8782**	**0.8942**	**0.8708**	0.7157	**0.7654**
MeDi5	0.9892	0.9334	0.8943	0.8812	0.8735	0.8648	0.8848	0.8590	0.7222	0.7452

* The MeDi1 can be regarded as the image preprocessed by median filtering.

Compared with the single noise in Section 4.4, the degradation model further increases the complexity of noise. The results in Tab. 1 and Tab. 2 demonstrate that MeDiNet can enhance the performance under real-world noisy condition. We can observe improvements of GB+AWGN, GB+SP, and GB+HG are significant. The performance under the GB+MG is still limited.

In order to fully verify the effectiveness of our method, we also report the results of CA-LFW and CP-LFW dataset (See Tab. 2). These datasets are collected for specific challenges (pose and age difference). We can observe that the improvement on the CP-LFW dataset is limited. This dataset selects face pairs with pose differences to add pose variation to intra-class variance and contains less edge information than LFW. When using the MeDiConv, the degradation of edge leads to a limited increase on the CP-LFW (e.g., the accuracy of MeDi5 is only 3.32% higher than baseline under GB+MG condition).

Table 2: The evaluation on the CA-LFW and CP-LFW dataset

	Clean		GB+AWGN		GB+SP		GB+HG		GB+MG	
	CP	CA	CP	CA	CP	CA	CP	CA	CP	CA
Sphere20	0.7618	**0.9078**	0.5697	0.6957	0.5238	0.5545	0.5502	0.6030	0.5130	0.5403
MeDi1	0.7607	0.9020	0.5868	0.7288	0.5525	0.6500	0.5500	0.6405	0.5133	0.5337
MeDi2	**0.7638**	0.9032	0.5927	0.7647	0.5677	0.6903	0.5675	0.6835	0.5315	0.5498
MeDi3	0.7620	0.9017	0.6028	0.7632	0.5660	0.6940	0.5707	0.6912	0.5405	0.5577
MeDi4	0.7407	0.8938	**0.6485**	**0.8097**	**0.6210**	0.7560	**0.6097**	0.7558	0.5372	0.6322
MeDi5	0.7553	0.8927	0.6438	0.8082	0.6132	**0.7628**	0.6028	**0.7582**	**0.5462**	**0.6327**

The evaluation of four datasets demonstrates that proposed MeDiNet enhances the performance of face recognition under the real-world noisy conditions. The accuracy is improved when the number of MeDiConv layers increases. The median representation causes a probability of missing edge information. Thus, the performance slightly drops compared with baseline when testing on clean images. When adding multiple combinations of noises, the face recognition accuracy increases significantly on four datasets compared with baseline due to the noise-robustness of MeDiNet. However, the improvement under complex

noise (GB+HG and GB+MG) is lower than other combinations. The ability of MeDiNet to deal with complex noises needs to be strengthened. Solving the problem of edge caused by median representation is also the focus of future work.

In order to further analyze the characteristics of MeDiConv, we report the time consumption of the model and transfer it to the ResNet architecture for evaluation.

Time Consumption In the experiment, we observed median operation is hard to speed up on GPU. When the number of MeDiConv layers increases, the training time consumption significantly increases (See in Fig. 6). It would be more efficient to implement from scratch at C/Cuda level. For training convergence, the difference between the loss curve and training accuracy of MeDi1-3 and baseline is limited, illustrating that training MeDiNet is stable.
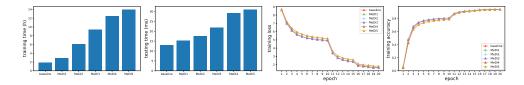

Figure 6: The first graph shows the training time of each model (20 epochs); The second graph shows the inference time for per image; The third and fourth graphs show the training loss and accuracy.

Transfer Ability For verifying that MeDiConv is a general way to handle noise, we choose ResNet18 as the backbone. The result is shown in Tab 3. We fixed the first convolution in ResNet as MeDiConv. The MeDiRes1 denotes that we replace the convolutional layers in the first block with MeDiConv. The result showed our methods still can suppress the negative impact of noise, which can prove that MeDiConv is a general way to handle noise that can be transferred to other architectures.

Table 3: The ResNet architecture evaluated on the LFW

	Clean	GB+AWGN	GB+SP	GB+HG	GB+MG
ResNet18	**0.9888**	0.8065	0.7000	0.7943	0.6665
MeDiRes1	0.9730	0.8348	0.7847	0.8205	0.6688
MeDiRes2	0.9810	**0.8580**	**0.8110**	**0.8378**	**0.6950**

5   Conclusion

We proposed a robust architecture MeDiNet, to enhance the performance of face recognition under noisy conditions. Standard CNNs are sensitive to the noise in the real world, especially in low-quality security cameras, low-light conditions, etc., which leads to a significant decrease in the accuracy of face recognition. Inspired by MRELBP descriptor, MeDiNet adopted median pixel difference convolutions, which can effectively remove noisy pixels in the feature map. Extensive experimental results showed the MeDiNet demonstrates striking robustness to various image noises.

Compared with standard CNNs, our method has attractive properties on noisy images. However, there are several challenges that need to be addressed: 1) The trade-off exists between accuracy and noisy robustness. The median representation can affect feature extraction (e.g., edge detection) in the deep layers; 2) The overlap region between neighbor MeDiConv kernel might result in the same median value, which can affect the accuracy of face recognition; 3) The accuracy still decreases significantly under the complex noise.

Acknowledgement

References

[1] A. Abdelhamed, S. Lin, and M.S. Brown. A high-quality denoising dataset for smart-phone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018.

[2] S. Anwar and N. Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3155–3164, 2019.

[3] E. Ataman, V. Aatre, and K. Wong. Some statistical properties of median filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(5):1073–1075, 1981.

[4] K. Cao, Y. Rong, C. Li, X. Tang, and C.C. Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2018.

[5] J. Chen, K. Li, Q. Deng, K. Li, and P.S. Yu. Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2019.

[6] G.G. Chrysos and S. Zafeiriou. Deep face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 69–78, 2017.

[7] Y. Ding, Y. Cheng, X. Cheng, and B. Li. Noise-resistant network: a deep-learning method for face recognition under noise. *EURASIP Journal on Image and Video Processing*, 2017:43, 2017.

[8] G. Guo and N. Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 2019.

[9] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I.W. Tsang, and M. Sugiyama. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *Proceedings of the 32th International Conference on Neural Information Processing Systems*, pages 8536–8546, 2018.

[10] W. Hu, Y. Huang, F. Zhang, and R. Li. Noise-tolerant paradigm for training face recognition cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11887–11896, 2019.

[11] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(6):765–781, 2011.

[12] G.B. Huang, M. Ramesh, T. Berg, and Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, 2007.

[13] F. Juefei-Xu, V. N. Boddeti, and M. Savvides. Local binary convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19–28, 2017.

[14] B. Kamgar-Parsi and W. Lawson. Toward development of a face recognition system for watchlist surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1925–1937, 2011.

[15] S. Kumawat, M. Verma, and S. Raman. Lbvcnn: Local binary volume convolutional neural network for facial expression recognition from image sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 0–0, 2019.

[16] B. Lahasan, S.L. Lutfi, and R. San-Segundo. A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression. *Artificial Intelligence Review*, 52:949–979, 2019.

[17] J. Li, K. Jin, N. Kubota, and Z. Ju. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350, 2020.

[18] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European Conference on Computer Vision*, pages 272–289, 2018.

[19] Y. Li, L. Song, X. Wu, R. He, and T. Tan. Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[20] Y. Lin, S. Cheng, J. Shen, and M. Pantic. Mobiface: A novel dataset for mobile face tracking in the wild. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8, 2019.

[21] L. Liu, S. Lao, P.W. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen. Median robust extended local binary pattern for texture classification. *IEEE Transactions on Image Processing*, 25(3):1368–1381, 2016.

[22] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen. From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127:74–109, 2018.

[23] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128:261–318, 2020.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[25] T. Ojala, M. Pietikäinen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[26] T. Plotz and S. Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1586–1595, 2017.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[28] Z. Su, M. Pietikäinen, and L. Liu. Bird: Learning binary and illumination robust descriptor for face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 102, 2019.

[29] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5117–5127, 2021.

[30] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 1988–1996, 2014.

[31] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[33] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S.J. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6575–6583, 2017.

[34] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2):1–37, 2013.

[35] S.K. Vipparthi, S. Murala, S.K. Nagar, K. Li, and A.B. Gonde. Local gabor maximum edge position octal patterns for image retrieval. *Neurocomputing*, 167(1):336–345, 2015.

[36] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, Qian C., and C.C. Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018.

[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 499–515, 2016.

[38] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011.

[39] X. Wu, M. Liu, Y. Cao, D. Ren, and W. Zuo. Unpaired learning of deep image denoising. In *Proceedings of the European Conference on Computer Vision*, pages 352–368, 2020.

[40] D. Yi, Z. Lei, and S.Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[41] X. Yu and F. Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4327–4333, 2017.

[42] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5295–5305, 2020.

[43] K. Zhang, W. Zhuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

[44] Y. Zhang, W. Deng, M. Wang, J. Hu, X. Li, D. Zhao, and D. Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7731–7740, 2020.

[45] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

[46] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.