Universal Face Restoration With Memorized Modulation

Jia Li, Huaibo Huang, Xiaofei Jia, Ran He
National Laboratory of Pattern Recognition, CASIA
Center for Excellence in Brain Science and Intelligence Technology, CAS
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract

Blind face restoration (BFR) is a challenging problem because of the uncertainty of the degradation patterns. This paper proposes a Restoration with Memorized Modulation (RMM) framework for universal BFR in diverse degraded scenes and heterogeneous domains. We apply random noise as well as unsupervised wavelet memory to adaptively modulate the face-enhancement generator, considering attentional denormalization in both layer and instance levels. Specifically, in the training stage, the low-level spatial feature embedding, the wavelet memory embedding obtained by wavelet transformation of the high-resolution image, as well as the disentangled high-level noise embeddings are integrated, with the guidance of attentional maps generated from layer normalization, instance normalization and the original feature map. These three embeddings are respectively associated with the spatial content, high-frequency texture details, and a learnable universal prior against other blind image degradation patterns. We store the spatial feature of the low-resolution image and the corresponding wavelet style code as key and value in the memory unit, respectively. In the test stage, the wavelet memory value whose corresponding spatial key is the most matching with that of the inferred image is retrieved to modulate the generator. Moreover, the universal prior learned from the random noise has been memorized by training the modulation network. Experimental results show the superiority of the proposed method compared with the state-of-the-art models, and a good generalization in the wild.

1. Introduction

Face super-resolution (FSR) is a challenging generation task, especially when it comes to large scale factors and unpaired real scenes. Generally, the low-quality (LQ) image contains the low-level spatial content information. The high-level texture details of the high-quality (HQ) image can be formulated as the high-frequency wavelet coefficients [49, 33, 18], which can potentially defend against the down-sampling degradation. Other degradation patterns in the real world are randomly comprised of Gaussian blurring, motion blurring, image noise, JPEG compaction, etc. This makes BFR more difficult than FSR. In this paper, we adversarially defend these complex degradations using noise as a universal learnable prior to modulate the blind face restoration. Comprehensively, we consider the representation learning of the degraded image, high-frequency wavelet prior, and an additional universal prior, which is different from the recent reference-based [31, 30, 9, 55, 34], HQ-dictionary-based [29] and pretrained-model-based [37, 5, 46] SR methods that have shown great potential to improve the restoration quality.

Memory networks were firstly proposed in [48] for the task of question answering where the memory acts as an effective knowledge base. As for the computer vision field, there are some applications, such as image captioning [8], text-to-image synthesis [56] and object segmentation [38]. We first propose to realize BFR with wavelet coefficient memory knowledge. We address SR task via attention-guided feature modulation of noise and wavelet memory, not relying on additional guidance or prior from HQ reference image [31, 10] in the inference stage, or pretrained well-performing models (e.g. StyleGAN [21, 22]) in the training stage [5, 46]. We propose a novel BFR model called RMM that synthesizes high fidelity high-resolution (HR) face results both for the synthetic and real-world LQ images. Specifically, RMM uses the low resolution (LR) face image to carry out adaptive spatial modulation of the multi-scale decoder features, and makes use of the disentangled multi-scale noise embeddings and high-frequency wavelet code to implement the BFR information modulation. Actually, the noise embedding is resolution-independent whose latent size is fixed and with Gaussian distribution, whereas wavelet memory embedding is resolution-dependent whose distribution of style code is related to the degradation degree, i.e., the larger the down-sampling scale is, the more feature dimensions wavelet embedding has.

Adaptive feature modulation has shown great benefit to
improving the desired style transferring for various synthesis tasks. [41] synthesizes photorealistic images using SPAtially-Adaptive DENormalization (SPADE) based on the semantic layout modulation. StyleGAN [21, 22] generates high-quality images via adaptive instance normalization (AdaIN) based on noise modulation. [28] combines SPADE of the target face and AdaIN of the source identity to address the high fidelity face swapping. Furthermore, [23] proposes Adaptive Layer-Instance Normalization (AdaLIN) to flexibly control the fidelity of shape and texture across the heterogeneous domains, where instance normalization (IN) [44] and layer normalization (LN) [1] are combined via a learnable parameter. [27] adopts adaptive point-wise layer instance normalization (AdaPoLin) by combining IN and LN via a convolutional layer in an all-channel manner. Different from the above adaptive modulation manner, in this paper, RMM pays attention to three-level Attentional maps from the Original, Layer-Normalized and Instance-Normalized features, denoted as AdaAO, AdaALN and AdaAIN. AdaAO integrates the feature maps from AdaALN and AdaAIN that adaptively fuse the LR feature, wavelet memory and noise prior with corresponding attentional maps in the layer and instance modulation view, respectively. This enables RMM to comprehensively control the structure and detail textures with consideration of the local and global statistics for the multi-level feature maps. Our main contributions are as follows:

• We first propose an effective and efficient Wavelet Memory Module (WMM) that stores spatial features of the LQ images and high-frequency wavelet coefficient information, which enables to adaptively guide the high-quality blind face restoration.

• We first propose an effective and efficient Wavelet Memory Module (WMM) that stores spatial features of the LQ images and high-frequency wavelet coefficient information, which enables to adaptively guide the high-quality blind face restoration.

2. Related Work

Single Image Restoration With the development of CNNs, GANs and flow-based model, image restoration from a single image has made great progress in many fields, including image super-resolution [12, 32, 54], deblurring [50, 25], decompression [11, 14], and denoising [52, 13]. The typical SR task without guidance is WaveletSR [18] where the predicted wavelet coefficients of HQ images are composed to reconstruct the restored face. [49] adopts an Invertible Rescaling Net (IRN) to model the dual mapping of LQ and HQ images. HiFaceGAN [51] uses a collaborative suppression and replenishment (CSR) framework to achieve face renovation.

Model-based Image Restoration GLEAN [5] utilizes pre-trained StyleGAN to provide more detail features with the distribution of HR images. GFP-GAN [46] is comprised of a degradation removal module and a pretrained StyleGAN2 [22] as the face prior. PULSE [37] looks for the optimized latent code of StyleGAN [21] by penalizing a down-scaling loss between the LR image and the degraded SR image.

Reference-based Image Restoration WarpNet [31] makes use of the degraded and guided images to produce the restoration result. [30] incorporates prior of the reference image where both degraded, guidance features and facial landmarks are used to calculate an attention mask for adaptive high-resolution feature fusion. [9] builds up a universal reference pool to match the local patterns for adaptive feature enhancement of the different regions. [55] leverages internal and external priors to recover the HR face image. [34] handles different reference images by remapping the distribution of guided features to that of the LR features by spatial adaptive transferring. PSFRGAN [6] proposes a multi-scale progressive framework for BFR based on the parsing prior. [29] applies dictionary feature transfer (DFT) block to adaptively match with the off-line multi-scale facial component dictionaries.

3. Approach

In this section, we will introduce our method RMM in detail, including the RM framework, objective function, and instance-normalized features, i.e., AdaAO, AdaALN and AdaAIN, to implement the quality improvement both visually and quantitatively.
wavelet memory module (WMM). As shown in Figure 2, given a target face \( X_{HR} \), we obtain the degraded image \( X_{LR} \) by means of the sequential degradation methods, e.g., downsampling with the bicubic kernel followed by adding blurring and JPEG compaction. \( X_{LR} \) is mapped to the refined middle-resolution face \( X_{MR} \) via an Unet [42] with only a reconstruction constraint, and we obtain multi-scale spatial feature maps \( z_S = \{z^1_S, z^2_S, ..., z^N_S\} \). We leverage a mapping network to disentangle the input random noise with Gaussian distribution, and utilize these prior embeddings \( \{z^1_N, z^2_N, ..., z^N_N\} \) to modulate multi-scale decoder of RMM. Another branch is the wavelet style code \( z_W \) generated from the WMM. These three embeddings are adaptively integrated by means of the \( RM^3 \) module.

\( RM^3 \) For the input feature \( h^i \) of each \( RM^3 \) block, it is modulated by three groups of affine transform parameters from \( z^1_S \), \( z^1_N \) and \( z_W \), respectively. It is formulated as:

\[
h^i_L = \frac{h^i - \mu^i_L}{\sigma^i_L},
\]

\[
\{S, N, W\}^i_L = \gamma^i_{\{S, N, W\}} \odot h^i_L + \beta^i_{\{S, N, W\}},
\]

where \( h^i \in \mathbb{R}^{C^i \times H^i \times W^i} \) is the input embedding of the current \( RM^3 \) block, \( \mu^i_L \) and \( \sigma^i_L \) are the means and standard deviations of \( h^i \), and they are used to perform the instance normalization. Let \( z^1_S \in \mathbb{R}^{C^S \times H^S \times W^S} \), \( z^1_N \in \mathbb{R}^{C^N \times 1} \) and \( z_W \in \mathbb{R}^{C_W \times 1} \) be the LR, wavelet and noise embedding, respectively. As shown in Figure 2, \( \gamma^i_{S} \) and \( \beta^i_{S} \) are obtained from \( z^1_S \) using a convolutional layer. Meanwhile, \( \{\gamma^i_{N}, \beta^i_{N}\} \in \mathbb{R}^{C^i \times H^i \times W^i} \) and \( \{\gamma^i_{W}, \beta^i_{W}\} \in \mathbb{R}^{C^i \times H^i \times W^i} \) are obtained from \( z^1_N \) and \( z_W \) using a fully connected layer, respectively. In the AdaAIN form, the attentional map \( M^i_{S}, M^i_{N}, \) and \( M^i_{W} \), are obtained via a convolution and a softmax operation of \( h^i_L \). The instance denormalization result \( H^i_L \) is formulated as

\[
H^i_L = S^i_L \odot M^i_{S} + N^i_L \odot M^i_{N} + W^i_L \odot M^i_{W},
\]

where \( \odot \) means the element-wise product. Similarly, the layer normalization and denormalization are as follows

\[
h^i_L = h^i - \mu^i_L \sigma^i_L^{-1},
\]

\[
\{S, N, W\}^i_L = \gamma^i_{\{S, N, W\}} h^i_L + \beta^i_{\{S, N, W\}},
\]

In the AdaALN form, the attentional map \( M^i_{S}, M^i_{N}, \) and \( M^i_{W} \), are obtained via a convolution and a softmax operation of \( h^i_L \). The layer denormalization result \( H^i_L \) is formulated as

\[
H^i_L = S^i_L \odot M^i_{S} + N^i_L \odot M^i_{N} + W^i_L \odot M^i_{W},
\]

Finally, in the AdaAO form, the attentional map \( M^i_{O} \) is obtained via a convolution and a sigmoid operation of \( h^i_L \). The \( RM^3 \) output feature \( H^i \) is formulated as

\[
H^i = H^i_L \odot M^i_{O} + H^i_L \odot (1 - M^i_{O}).
\]

**Objective Function** Our BFR network RMM is trained with the following constraints.

We apply the multi-scale adversarial learning [45]. Let \( \mathcal{L}_{adv} \) be the adversarial loss to discriminate the generated BFR image \( \hat{X}_{HR} \) and the real HR face \( X_{HR} \) via

\[
\mathcal{L}_{adv} = \lambda_{adv} \sum_{i=1}^{N} E_{X} [\log D_{i}(X_{HR})] + E_{Y} [\log(1 - D_{i}(\hat{X}_{HR}))],
\]

Figure 2: The framework of RMM. The spatial features are used to maintain the global structure during BFR, and the resolution-dependent component wavelet style code from the high-frequency wavelet coefficients controls the feature transformation of the texture details. Moreover, the resolution-independent component Gaussian noise learns a universal prior to defend against diverse real-world degradation patterns. These BFR embeddings are adaptively integrated by \( RM^3 \) module, where different kinds of attentional maps from the original, layer-normalized and instance-normalized features are taken into consideration for a comprehensive feature fusion.
where $D_i$ means the discriminator for the downsampled image with scale factor $i$.

We utilize two reconstruction losses used to penalize the pixel-level distance between $X_{HR}$ and $\hat{X}_{HR}$, similar to $X_{HR}$ and $X_{MR}$. They are formulated as

$$L_{rec} = \begin{cases} \frac{1}{2} \Delta^2 \hat{X}_{HR} - \frac{1}{2} \delta & \delta 
 \end{cases},$$  

$$L_{rec}' = \begin{cases} \frac{1}{2} \Delta^2 \hat{X}_{MR} - \frac{1}{2} \delta & \delta 
 \end{cases},$$

respectively. We use Huber loss, where $\Delta \hat{X}_{HR} = \left| X_{HR} - \hat{X}_{HR} \right|$, $\Delta X_{MR} = \left| X_{HR} - X_{MR} \right|$, and $\delta$ is a hyperparameter.

We add the perceptual loss to improve the high-resolution blind face restoration via

$$L_{vgg} = \lambda_{vgg} \sum_{i=1}^{N} \left\| F_{vgg}^{i}(\hat{X}_{HR}) - F_{vgg}^{i}(X_{HR}) \right\|_2,$$

where $F_{vgg}^{i}$ denotes the $i$-th convolution layer of the VGG19 model. We set $N$ equal to 5. $\lambda_{vgg}$ is set to 1/32, 1/16, 1/8, 1/4 and 1.0 in order.

Apart from the above mentioned global objective functions, we further propose to utilize the component contextual loss to enhance the eyes and mouth areas via

$$L_{cCX} = -\log(CX(F_c^{i}(\hat{X}_{HR}), F_c^{i}(X_{HR}))),$$

where $l \in \text{relu}\{2, 4, 2\}$ layers of the pretrained VGG19 network [43], and $c$ means {left_eye, right_eye, mouth}. Contextual loss [36] is competent to learning the non-aligned feature matching. Finally, the total loss of RMM is formulated as:

$$L_{RMM} = L_{adv} + L_{rec}' + \lambda_{rec} L_{rec} + L_{vgg} + \lambda_{cCX} L_{cCX}.$$

**WMM** We apply an augment network Wavelet Memory Module to store the LQ spatial feature and high-frequency wavelet coefficients. As each Wavelet Memory Unit $WMU_i$, it contains the spatial feature key $K_S[t_i]$ and the corresponding wavelet code value $V_W[t_i]$, where $t_i$ means the top-$i$ value after K Nearest Neighbors operation between the key candidates and the input query. This memory mining is formulated as

$$KNN(q, K_S) = \arg \max_i \langle q, K_S[i] \rangle,$$

where $\langle \cdot, \cdot \rangle$ means cosine similarity. We use pretrained ResNet18 [15] as the spatial feature extractor, and generate the normalized query $q$ after transforming the spatial code from pool5 layer using a fully connected layer.

High-frequency wavelet coefficients help to recover texture details. Consider $n$-level full wavelet packet decomposition, where $n$ means the down-scaling factor of the original image. Then the wavelet coefficient is represented as $C = \{c_1, c_2, \ldots, c_{N_W} \}$, where $N_W = 4^n$, and $c_1$ is the low-frequency wavelet coefficient. We treat $C_h \in \mathbb{R}^{(N_W-1) \times H^{CH} \times W^{CH}}$ as the high-level feature map and obtain the wavelet memory code $Z_W \in \mathbb{R}^{(N_W-1) \times 1}$ by means of the Adaptive Average pooling for $C_h$. We use Wavelet KL divergence ($KL_{w}$) to guide the triplet loss of WMM in an unsupervised manner via

$$L_{WMM} = \max(\langle q, K_S[t_p] \rangle - \langle q, K_S[t_n] \rangle + m, 0),$$

where $m$ is the margin. Specifically, we minimize the distance between the positive key $K_S[t_p]$ and the query $q$ when $KL_w(V_W[t_p] \| Z_W) < \eta$, while maximizing the distance between the negative key $K_S[t_n]$ and the query $q$ when $KL_w(V_W[t_n] \| Z_W) > \eta$, where $\eta$ is a hyperparameter. The top-1 key $K_S[t_1]$ is updated as the mean of $q$ and the previous $K_S[t_1]$ when $KL_w(V_W[t_1] \| Z_W) < \eta$, while the new key $q$ and value $Z_W(q)$ will be written into wavelet memory as $K_S[t_1]$ and $V_W[t_1]$ when $KL_w(V_W[t_1] \| Z_W) > \eta$.

## 4. Experiment

### 4.1. Experimental Protocol

**Training Protocol** We train our RMM on the FFHQ [21] dataset with 512 size. The training batch size is set to 8. Each facial component {left_eye, right_eye, mouth} is cropped based on the face landmarks [3]. We train our model with Adam optimizer [24], and set $\beta_1 = 0.5$, $\beta_2 = 0.999$, learning rate $lr = 2e^{-4}$. We employ $D_i$ with $i = \{1, 2, 4, 8\}$, and the corresponding weights are 4, 2, 1, 1. We set $\lambda_{rec} = 100$, $\lambda_{cCX} = 1$. Moreover, the margin of $L_{WMM}$ is set to 0.1, the threshold $\eta$ is set to 0.7. The memory size of WMM is set to 982.

To synthetize the training data that possesses the similar distribution to the real degraded images, we apply the degradation model adopted in [31, 29, 46],

$$X_{LR} = (X_{HR} \otimes \{k_g, k_M\})_{lr} + n_{\sigma},$$

where $k_g$ denotes Gaussian blur kernel with $G \in \{1 : 5\}$, $k_M$ denotes the motion blur kernels [26]. Moreover, the down-sampling scale $r$, Gaussian noise $n_{\sigma}$, and JPEG compression quality $q$ are in the range of $\{2 : 12\}$, $\{1 : 15\}$, and $\{40 : 80\}$, respectively. We randomly integrate these degradation operations in the training stage.
Test Datasets We test the performance of our RMM with other state-of-the-art methods based on four synthetic datasets and two different real-world datasets as follows.

- **CelebA-Test-Blind** is the synthetic dataset with 2,000 CelebA-HQ images from its test set [20]. The generation way follows the degradation protocol in our training stage. This item is denoted as CelebA-TB.

- **CelebA-Test-Downsampling** is the original 2,000 CelebA-HQ images for the test, which is obtained directly by downsampling 4, 8, or 16 times. This item is denoted as CelebA-TD.

- **VGGFace2-Test-Blind** is the synthetic dataset with 3,000 test images of VggFace2 [4]. Similarly, it is denoted as VGGFace2-TB.

- **VGGFace2-Test-Downsampling** is denoted as VGGFace2-TD.

- **CFW-Test.** IIIT-CFW [39] is used in the cartoon face classification and Photo2Cartoon tasks in the wild. We select 252 challenging LQ images with resolution lower than $80 \times 80$.

- **LFW-Test.** [17] contains LQ images in the wild. We collect 1711 testing images with distinct identities in the validation partition, where the image with the first index is selected.

4.2. Comparisons with State-of-the-art Methods

For BFR task, we employ several evaluation metrics considering the image fidelity (FID [16], KID [2], NIQE [40]), pixel-wise perception (PSNR, SSIM, MS-SSIM), patch-wise perception (LPIPS [53]) and the identity preservation (FED, LLE). Specifically, KID [2] is the Kernel Inception Distance using the squared Maximum Mean Discrepancy (MMD) with a polynomial kernel, which is used in [23, 7]. NIQE [40] is a non-reference blind image quality assessment for grayscale images, MS-SSIM considers multi-scale SSIM between the FSR result and the ground truth. Moreover, we measure the feature embedding distance (FED) and landmark localization error (LLE) with a pretrained face recognition model based on dlib toolkit. We evaluate FID and KID by calculating the distribution distances between the BFR results and the real images from FFHQ...
CelebA-TB, CFW-Test, and LFW-Test are considered. The results of RMM (col 9) are with higher w/o Z spatial feature and wavelet memory to modulate RMM, and Z w/o Mapping 4.3. Ablation Study

5 for the large face pose, and Figure 7 for the old photos of and background in col 6). More results are shown in Figure 8. This will cause a serious identity distortion, as shown in col 5, Figure 3. Our RMM has more competitive performance under comprehensive evaluation metrics. Therefore, the restored faces are higher-fidelity than that of w/o Z_N and w/o Z_W, and can capture more precise texture and structural details, e.g., the shape of eyes, mouth and tooth.

Whether depending on X_MR As we all know, X_MR is supposed to refine the global spatial information. We conduct more ablation studies, as shown in Figure 6.

(a). If directly using the input degraded spatial feature from X_LR, denoted as w/o X_MR, there will be lots of artifacts, especially in the seriously distortion cases.

(b). If only with X_MR, i.e., there is no wavelet memory and universal prior modulations, the restored results are messy as well, which demonstrates the essential role of RM^3 module.

(c). While w/X_MR and w/Z_W, it is still challenging for the model to defend the degradation, and some facial components, e.g., eyes or nose, can not be restored correctly.

(d). While w/X_MR and w/Z_N, although the corresponding results are reasonable and smooth, they lack some high-frequency details, e.g., the subtle contour of tooth (Figure 4), eyes & lip (Figure 6).

(e). While using the full framework, i.e., RMM, the results are considerable with respect to the global topology and high-frequency details, which is more robust in the wild and heterogeneous domains.

Wavelet Memory Bank

Spatial wavelet modulation The reason of using the latent code not the spatial wavelet feature is to prevent the overfitting. Specifically, the spatial wavelet memory is very helpful to the restoration in the training stage, but causes a bad impact on the samples in the test stage. No two leaves in the world are exactly the same, and the faces are as well. No two leaves in the world are exactly the same, and the faces are as well.

4.3. Ablation Study

The qualitative results are shown in Figure 4. w/o Mapping means discarding the mapping network for the input noise Z. w/o X_MR means no refinement of the spatial feature for the input LQ image. w/o AdaALN and w/o AdaAIN mean only AdaAIN or AdaALN is conducted, respectively. w/o Z_N denotes that we utilize the spatial feature and wavelet memory to modulate RMM, and w/o Z_W means only noise prior and spatial modulations are considered. The results of RMM (col 9) are with higher fidelity than that of w/o Z_N and w/o Z_W, and can capture more precise texture and structural details, e.g., the shape of eyes, mouth and tooth.

Whether depending on X_MR As we all know, X_MR is supposed to refine the global spatial information. We conduct more ablation studies, as shown in Figure 6.

(a). If directly using the input degraded spatial feature from X_LR, denoted as w/o X_MR, there will be lots of artifacts, especially in the seriously distortion cases.

(b). If only with X_MR, i.e., there is no wavelet memory and universal prior modulations, the restored results are messy as well, which demonstrates the essential role of RM^3 module.

(c). While w/X_MR and w/Z_W, it is still challenging for the model to defend the degradation, and some facial components, e.g., eyes or nose, can not be restored correctly.

(d). While w/X_MR and w/Z_N, although the corresponding results are reasonable and smooth, they lack some high-frequency details, e.g., the subtle contour of tooth (Figure 4), eyes & lip (Figure 6).

(e). While using the full framework, i.e., RMM, the results are considerable with respect to the global topology and high-frequency details, which is more robust in the wild and heterogeneous domains.
Figure 6: The ablation results in the seriously distortion situations.

Figure 7: More detailed faces in old photos.

Figure 8: The comparison results with original wavelet feature modulation, which obtained not through average pooling.

Figure 9: The results of CNN feature modulation based on VGG19.

Table 2: Quantitative evaluation on RMM variants with different memory sizes.

| Dataset | Mean |
|---------|------|
| Mem_10  | 102.1 8.8 4.5343 |
| Mem_50  | 102.45 8.83 4.546 |
| Mem_100 | 99.78 8.42 4.7282 |
| Mem_200 | 102.61 8.73 4.5686 |
| Mem_500 | 104.76 8.84 4.7031 |
| Mem_1000| 100.38 8.48 4.5309 |

Table 3: Quantitative evaluation on RMM variants with WPD from different feature layers.

| Dataset | Mean |
|---------|------|
| RMMConv1 | 110.41 9.29 4.29 |
| RMMConv2 | 126.04 11.27 5.90 |
| RMMConv3 | 125.54 11.13 6.46 |
| RMM    | 100.61 8.24 4.54 |

Memory size We conduct the ablation for RMM variants with different memory sizes. Results show that FID, KID and NIQE scores are stable across a wide range of wavelet memory sizes, as shown in Table 2. Mem_100 has the best FID and KID scores, but with a high NIQE score. Whereas Mem_1000 has a competitive performance of FID and KID, and has the best NIQE score. Furthermore, the larger memory size can contain more wavelet patterns, which is benefit to the model generalization.

WPD of shallow layer We use wavelet packet decomposition of features from shallow layers, i.e., conv1_1, conv2_1 and conv3_1 of VGG19, to modulate RMM. As shown in the bellow table, RMMConv1 has the best NIQE score, but with worse FID and KID, compared with RMM. With WPD of deeper layer, the restored image is more blurred and has bad performance on FID, KID and NIQE scores. On the contrary, RMM uses the WPD of the input image, the structural and textural details are stored to a great extent, to conduct an effective high-frequency feature modulation.

Query The query vector is associated with the low-quality image. We use features from shallow layers as the query, e.g., ResNet18 subsequences with index 0, 2, 4 and 6 are used to produce the query, respectively. Note that our
Figure 10: Comparison results of WPD for the shallow layers and the input image. RMM directly uses the WPD of the input face.

Figure 11: Comparison results for query from different shallow layers.

Table 4: Quantitative evaluation on RMM variants with query from different shallow layers.

| Dataset | Mean FID | Mean KID | Mean NIQE |
|---------|---------|----------|-----------|
| Resnet | 106.9 | 9.11 | 4.27 |
| Resnet | 106.04 | 8.64 | 4.38 |
| Resnet | 106.1 | 8.68 | 4.24 |
| Resnet | 107.33 | 8.72 | 4.46 |
| RMM | 106.61 | 8.24 | 4.54 |

RMM adopts the eighth model sequence followed by adaptive average pooling. These settings have higher FID and KID scores, although with slightly lower NIQE scores.

As for the visual experiment, we find more artifacts, e.g., wrong and messy details, are synthesized in the eyes, mouth, or background areas. The reason may be that the perception of the query generated from a deep layer, e.g., pool5 layer, is more global and comprehensive, so the retrieved high-frequency wavelet features are more accurate and matched with the facial content. However, the query of shallow layers is not capable of representing the spatial feature of LQ image well. Because there are lots of low-level degradations in the LQ image, the spatial query from shallow layers may have more mistakes.

4.4. Attentional Map Visualization

We show the multi-level attentional maps of different decoder features with RMM modulation for the inferred LQ face, in Figure 12. The bright area means the activation for different feature embeddings. $M^i_{D}$ shows that RMM pays more attention on layer denormalization (LDN) in the lower-level layers (rows 2&3&7), while the middle layers (row 4&5) have more attentions on the instance denormalization (IDN). Face edges (row 6) are also activated in the IDN. The proposed 7 kinds of attentional maps are meaningful for the comprehensive feature fusion. Although there are few activations of $M^i_{N_J}$ and $M^i_{N_L}$ for LDN and IDN of the test images, the noise modulation has an essential role for RMM training, as shown in Figure 2 (right) and Figure 4 (w/o $Z_W$, w/o $Z_N$).

4.5. Failure cases

There are two hard cases. One is for the generalization in heterogeneous domains, if the cartoon face is similar to the real-world face, e.g., the eye size is closed to that of humans, the restored eyes will lose the cartoon style. The other is the face with strange makeup, which makes it so ambiguous that the facial component is wrongly restored.

Figure 12: Multi-level attentional maps of RMM.

Figure 13: Some failure cases of RMM in the heterogeneous domain. RMM respects the original structural and color distribution of the LQ face.

5. Conclusion

We propose an RMM framework for blind face restoration. We apply random noise as well as unsupervised wavelet memory to adaptively modulate the BFR generator, considering AdaAO, AdaALN and AdaAIN. RMM controls three embedding modulations respectively associated with the global spatial content, high-frequency texture and structural details, as well as a learnable universal prior against diverse blind image degradation patterns. Experimental results show the superiority of the memorized modulation method on blind face enhancement and a good generalization in the wild and heterogeneous domains, e.g., oil painting, 3D cartoons, pencil drawing, exaggerated drawing and NIR image.
References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 2
[2] Mikolaj Binkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In ICLR, 2018. 5
[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In ICCV, 2017. 4
[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and An- derw Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pages 67–74. IEEE, 2018. 5, 6, 14
[5] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. arXiv preprint arXiv:2012.00739, 2020. 1, 2, 6
[6] Chao Dong, Chen Change Loy, and Andrew Zisserman. Progressive semantic-aware style transformation for blind face restoration. arXiv preprint arXiv:2009.08709, 2020. 2, 5, 6
[7] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In CVPR, pages 8168–8177, 2020. 5
[8] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In CVPR, pages 895–903, 2017. 1
[9] Shuguang Cui. Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation. 2020. 2020. 1, 2
[10] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In CVPR Workshops, pages 0–0, 2019. 1
[11] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaouq Tang. Compression artifacts reduction by a deep convolutional network. In ICCV, pages 576–584, 2015. 2
[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaouq Tang. Learning a deep convolutional network for image super-resolution. In ECCV, pages 184–199. Springer, 2014. 2
[13] Majed El Helou, Ruofan Zhou, and Sabine Süsstrunk. Stochastic frequency masking to improve super-resolution and denoising networks. In ECCV, pages 749–766. Springer, 2020. 2
[14] Jun Guo and Hongyang Chao. Building dual-domain representations for compression artifacts reduction. In ECCV, pages 628–644. Springer, 2016. 2
[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 4
[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NIPS, 2017. 5
[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Workshop on faces in ‘Real-Life’ Images: detection, alignment, and recognition, 2008. 5
[18] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-snet: A wavelet-based cnn for multi-scale face super resolution. In ICCV, pages 1689–1697, 2017. 1, 2, 14
[19] Janspiry. Color similarity calculation. https://github.com/Janspiry/Color-Similarity-Calculation. 11
[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017. 5, 6, 14
[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In CVPR, pages 4401–4410. 2019. 1, 2, 4, 6, 14
[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In CVPR, pages 8110–8119, 2020. 1, 2
[23] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In ICLR, 2020. 2, 5
[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 4
[25] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In CVPR, pages 8183–8192, 2018. 2
[26] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In CVPR, pages 1964–1971. IEEE, 2009. 4
[27] Bing Li, Yuanlue Zhu, Yitong Wang, Chia-Wen Lin, Bernard Ghanem, and Linlin Shen. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation. arXiv preprint arXiv:2102.12593, 2021. 2
[28] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In CVPR, June 2020. 2
[29] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In ECCV, pages 399–415. Springer. 2020. 1, 2, 4, 5, 6, 14
[30] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In CVPR, pages 2706–2715, 2020. 1, 2
[31] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In ECCV, pages 272–289, 2018. 1, 2, 4
[32] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In NIPS, volume 31. Curran Associates, Inc., 2018. 2
[33] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In CVPR workshops, pages 773–782, 2018.

[34] Xin Tao, Jianglei Zhu, Jiayu Lai, and Wenbo Li. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In CVPR, 2021.

[35] Lotayou. Metrics_package. https://github.com/Lotayou/Face-Renovation.

[36] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In ECCV, pages 768–783, 2018.

[37] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In CVPR, pages 2437–2445, 2020.

[38] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In CVPR, pages 10366–10375, 2020.

[39] Ashutosh Mishra, Alexandre Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In CVPR, pages 2437–2445, 2020.

[40] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. IEEE Signal processing letters, 20(3):209–212, 2012.

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In CVPR, pages 2337–2346, 2019.

[42] 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.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016.

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In CVPR, pages 8798–8807, 2018.

[46] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. arXiv preprint arXiv:2101.04061, 2021.

[47] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[48] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In ICLR, 2015.

[49] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In ECCV, pages 126–144. Springer, 2020.

[50] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. NIPS, 27:1790–1798, 2014.

[51] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifarcegan: Face renovation via collaborative suppression and replenishment. In ACM MM, pages 1551–1560, 2020.

[52] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing, 26(7):3142–3155, 2017.

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, pages 586–595, 2018.

[54] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In ECCV, pages 286–301, 2018.

[55] Yang Zhang, Ivo W Tsang, Yawei Luo, Chang-Hui Hu, Xiaobo Lu, and Xin Yu. Copy and paste gan: Face hallucination from shaded thumbnails. In CVPR, pages 7355–7364, 2020.

[56] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dmgan: Dynamic memory generative adversarial networks for text-to-image synthesis. In CVPR, pages 5802–5810, 2019.
Supplementary material

Additional Experimental Results

Motion blur  Figure 14 shows the test images with motion blur, which demonstrates that our RMM can handle casual motion blur degradation, even in the heterogeneous domain.

Noise degradation factor  Figure 15 shows the test images with noise degradation, which demonstrates that our RMM can handle random noise degradation, even in the heterogeneous domain.

GFP-GAN  We compare with the sota method GFP-GAN. Although our RMM does not depend on the pretrained StyleGAN prior, our performance is also competitive. We show more comparison results for the heterogeneous domains (oil painting, 3D cartoons, pencil drawing and exaggerated drawing) in Figure 17.

It should be noted that our results preserve the image color well, whereas the restored faces of GFP-GAN have a larger color gap to the original LQ faces, as shown in Table 5.

Table 5: Color similarity scores of GFP-GAN and RMM based on Hist-CIEDE [19]. The smaller score is the best, which indicates the restored face has a higher color fidelity of the LQ image.

| Dataset | CelebA-HR | Vist4002-HR | LFW-test | CPW-test |
|---------|-----------|-------------|----------|----------|
| GFP-GAN | 7.97      | 8.31        | 12.31    | 17.96    |
| RMM     | 7.51      | 8.30        | 12.34    | 17.96    |

Color consistency  Coloring the grayscale image may not be the drawback of GFP-GAN, which is considerable in the photographic domain. However, it is a fact that there is a larger color gap between the restored and original LQ image for GFP-GAN. Especially it lacks controllability in the heterogeneous domains, e.g., the color information of the colored face and the original background will be inconsistent in the sketch or grayscale domain. When it comes to another challenging case, e.g., NIR image enhancement, the lighting information on the face is supposed to be maintained. Otherwise, the scene presentation looks unharmonious, e.g., the NIR portrait seems daytime VIS portrait, as shown in Figure 19.

On one hand, the colored restored face may look more pleasing. On the other hand, GFP-GAN does not respect the original color attribute well. Our RMM has a competitive performance compared with GFP-GAN, and better color controllability. Moreover, RMM has a better generalization in the wild. There are more accurate structural and textural facial details, but fewer artifacts, thanks to the effectiveness of the wavelet memory and the universal prior, as shown in Figure 20.

Generalization in heterogeneous domains  There are three important aspects for the generalization in heterogeneous domains. The first is the preservation of the global structural and color information from the LQ image, the second is the accurate high-frequency textual restoration using the memorized wavelet modulation, and the last is the ability to defend diverse blind degradations using the learned universal restoration prior.

Because of these functions, our RMM respects the global attributes of the original image (e.g., in the sketch, grayscale, or oil drawing domain), complements accurate high-frequency local details (e.g., in the 3D cartoon or exaggerated drawing domain), and improves the model robustness and controllability in the wild. RMM realizes reasonable and high-fidelity facial restoration both in the photographic and heterogeneous domains.

Additional Analysis of RMM  We set the dimension of \( Z_W \) and \( Z_N \) to 765 and 512, respectively. If randomly sampling different \( Z_N \) or choosing different top-\( i \) wavelet memory \( Z_W \) to implement the modulation in the inference stage, the visual results maintain the same. However, the residual maps in Figure 18 (up) demonstrate that the low-rank wavelet memory has more mistakes of the high-frequency coefficient, e.g., the eye area of \( X_{SR}^{wm_{top}} - X_{SR}^{SR} \), while using the same \( Z_N \). Furthermore, we use the same top-\( i \) wavelet memory but different noises as the universal prior, and show the results in Figure 18 (below). We find that the global spatial information of the residual maps is maintained, although different noises focus on varied image areas for BFR.

Motivation of embedding prior  We think the key challenge of Blind face restoration is the restoration controllability and generalization. Our main aim is to restore the high-fidelity face with respect to the original structure and textual distribution, and improve the generalization in the wild and heterogeneous domains. To address this issue, we propose an effective and neat framework RMM. Specifically, the spatial feature \( Z_S \) helps to preserve the scene content and face identity by maintaining the global topology information. The wavelet style code \( Z_W \) helps to restore the high-frequency textual details by matching the memory knowledge. Note that wavelet packet decomposition (WPD) only considers the image magnification, and the random Gaussian noise in the original image remains in the WPD image as well, as shown in Figure 22. Based on the strong representation of wavelet coefficients, WaveletSR[3] has studied the face super-resolution task. However, BFR
Figure 14: More results on CelebA and the heterogeneous domain (the last) considering the motion blur. **Zoom in for the best view.**

Figure 15: More results on CelebA and the heterogeneous domain (the last) considering the random noise. **Zoom in for the best view.**

Figure 16: Some restored examples of GFP-GAN and RMM. **Zoom in for the best view.**

Figure 17: More examples of GFP-GAN and RMM in the wild, including the oil painting, 3D cartoons, pencil drawing and exaggerated drawing. **Zoom in for the best view.**

Figure 18: Additional analysis of RMM. The up and below are the results using different wavelet memories, and noises, respectively. Note that \( \text{wm}_{100} \) means RMM with top-100 wavelet memory, and \( \text{noise}_a \) means one randomly sampled Gaussian noise. **Zoom in for the best view.**

The task is more challenging than FSR task. In the real wild, the imaging and storage of image is easily interfered, which is very challenging for the model generalization. Therefore, the proposed universal prior \( Z_N \) is used to improve the model robustness by defending additional blind degradation pattern, such as motion blur, Gaussian noise, JPEG compression. Extensive experiments in the manuscript have shown that the proposed modulations of global spatial, universal prior and memorized wavelet embedding are beneficial and indispensable to restoring high-fidelity face in the complex scenes, and generalizing in the wild.

**RMM block size** We conduct the experiment considering the number of \( RM^3 \) blocks, and only with \( X_{MR} \), i.e., \( RM^3 \_0 \), as shown in Table 6. We apply three widely used
Figure 19: Comparison results in the NIR portrait.

Figure 20: More comparison results in heterogeneous domains, where the upper is in the exaggerated drawing domain, and the below is in the sketch domain.

Figure 21: The comparison results of DFDNet, GLEAN and RMM. Zoom in for the best view.

Figure 22: 1-level and 2-level full wavelet packet decomposition (WPD) image. The random Gaussian noise in the original image remains in the WPD image as well. Zoom in for the best view.

Figure 23: More examples only with $X_{MR}$. Zoom in for the best view.

test set, and obtain the mean scores. $RM^3_{-7}$ has the best scores, which indicates the progressive framework is effective, as shown in Figure 24.

Table 6: Quantitative evaluation on RMM variants with different $RM^3$ blocks.

| Dataset   | Mean     |
|-----------|----------|
| Methods   | FID↓    | KID↓    | NIQE↓  |
| $RM^3_{-7}$ | 111.46  | 10.52   | 4.7371 |
| $RM^3_{-1}$ | 104.16  | 10.63   | 4.5473 |
| $RM^3_{-2}$ | 107.51  | 8.80    | 4.5660 |
| $RM^3_{-3}$ | 103.10  | 8.36    | 4.6206 |
| $RM^3_{-4}$ | 106.20  | 9.00    | 4.5533 |
| $RM^3_{-5}$ | 101.29  | 8.29    | 4.5789 |
| $RM^3_{-6}$ | 100.81  | 8.36    | 4.5897 |
| $RM^3_{-7}$ | 100.61  | 8.24    | 4.5415 |

Figure 24: The results corresponding to RMM variants with different $RM^3$ blocks. Zoom in for the best view.

Figure 25: The right is the result of RMM applied to the movie *Amelie* (2001). Zoom in for the best view. Particularly, the eyes and lower eyelashes of the right are higher-quality.
Face Super Resolution

We compare our RMM with WaveletSR [18] and DFDNet [29] on the CelebA-TD and VGGFace2-TD, as shown in Table 7. The perceptual fidelity of FIDHQ is higher than [20, 4], and we get better quantitative results, e.g., FID, KID, compared with other state-of-the-art methods. DFDNet [29] is not competent to dealing with the LQ images with 16 × downsampling degradation, as shown in Figure 26. Note that WaveletSR [18] predicts high-frequency wavelet coefficients, so the pixel-wise perception and identity preservation metrics are better in 4 × inference setting. However, the image fidelity of WaveletSR [18], e.g., FID, KID, or NIQE, is not competitive. Note that the original test image of VGGFace2 is usually not high-quality, whereas our results are high-fidelity, so FID-VGGFace2 of our RMM is higher. More details are shown in Table 7.

Qualitative Results of RMM We show more results in Figure 27 on CelebA-TB, VGGFace2-TB, CFW-Test and LFW-Test, respectively. Moreover, we provide a demo of RMM in the supplementary material, and the video clip is from Amelie (2001). The original resolution is set to 1424 × 814, and the restored frame is 2848 × 1628. We show the comparison of a frame in Figure 25. For the stability of the background, we paste the restored face to the original frame using an open-source face parsing model.

Table 7: Quantitative evaluation on CelebA-TD and VGGFace2-TD. We train WaveletSR [18] based on FFHQ [21] dataset, CelebA-TD (row 1) and VGGFace2-TD (row 2).

| Methods | LPIPS (FFHQ/CelebA) | MS-SSIM | PSNR | SSIM | FID | KID | LLE | MS-SSIM | PSNR | SSIM | FID | KID | MS-SSIM | PSNR | SSIM | FID | KID | LLE |
|---------|---------------------|---------|------|------|-----|-----|-----|---------|------|------|-----|-----|---------|------|------|-----|-----|-----|
| Bicubic | 0.53                | 0.25    | 0.70 | 1.67 | 1.31| 0.73| 0.34| 0.80    | 0.20 | 0.22 | 1.00| 0.22| 0.80    | 0.20 | 0.22 | 1.00| 0.22| 1.00|
| WaveletSR [18] | 0.15 | 0.10 | 0.20 | 0.10 | 0.09 | 0.17 | 0.19 | 0.22 | 0.22 | 0.22 | 1.00| 0.22 | 1.00 | 0.22 | 1.00|
| DFDNet [29] | 0.08 | 0.06 | 0.10 | 0.06 | 0.05 | 0.15 | 0.17 | 0.19 | 0.22 | 0.22 | 1.00| 0.22 | 1.00 | 0.22 | 1.00|

Figure 26: Qualitative comparison with WaveletSR [18] and DFDNet [29] for FSR of different down-sampling scales on CelebA-TD (row 1) and VGGFace2-TD (row 2).
Figure 27: More results on CelebA-TB (left-up), VGGFace2-TB (right-up), CFW-Test (left-below) and LFW-Test (right-below). *Zoom in for the best view.*