Exploring Correlations in Degraded Spatial Identity Features for Blind Face Restoration

Qian Ning
ningqian@stu.xidian.edu.cn
School of Artificial Intelligence, Xidian University
China

Fangfang Wu*
wufangfang@xidian.edu.cn
School of Computer Science and Technology, Xidian University
China

Weisheng Dong
wsdong@mail.xidian.edu.cn
School of Artificial Intelligence, Xidian University
China

Xin Li
xli48@albany.edu
Department of Computer Science, University at Albany
USA

Guangming Shi
gmshi@xidian.edu.cn
School of Artificial Intelligence, Xidian University
China

ABSTRACT

Blind face restoration aims to recover high-quality face images from low-quality ones with complex and unknown degradation. Existing approaches have achieved promising performance by leveraging pre-trained dictionaries or generative priors. However, these methods may fail to exploit the full potential of degraded inputs and facial identity features due to complex degradation. To address this issue, we propose a novel method that explores the correlation of degraded spatial identity features by learning a general representation using memory network. Specifically, our approach enhances degraded features with more identity by leveraging similar facial features retrieved from memory network. We also propose a fusion approach that fuses memorized spatial features with GAN prior features via affine transformation and blending fusion to improve fidelity and realism. Additionally, the memory network is updated online in an unsupervised manner along with other modules, which obviates the requirement for pre-training. Experimental results on synthetic and popular real-world datasets demonstrate the effectiveness of our proposed method, which achieves at least comparable and often better performance than other state-of-the-art approaches.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; *Reconstruction.*

KEYWORDS

Face restoration, memory network, feature fusion, generative prior, StyleGAN

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611782

ACM Reference Format:
Qian Ning, Fangfang Wu, Weisheng Dong, Xin Li, and Guangming Shi. 2023. Exploring Correlations in Degraded Spatial Identity Features for Blind Face Restoration. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611782

1 INTRODUCTION

Blind face restoration (BFR) aims to recover high-quality (HQ) face images from low-quality (LQ) ones that are degraded by various factors, such as blur [30], low-resolution [4, 21, 25], compression artifacts [5, 24], and noise [6], among others. To tackle the challenge of BFR in real-world scenarios, many existing works [2, 19, 26, 32, 38] are trained on synthetic datasets featuring various degradations. To restore HQ face images with realistic details from degraded inputs, several facial-related priors have been adopted, such as facial parsing maps [3], facial component heatmaps [39], and 2D/3D facial landmarks [4, 11, 11]. Additionally, pre-trained dictionaries or codebooks [19, 34, 41, 42] have been leveraged to enhance the restoration quality. DFDNet [19] proposed to learn multi-scale dictionaries of eyes, noses, and mouths for blind face restoration. By combining degraded spatial identity features with previously learned dictionaries using Spatial Feature Transformer (SFT) [14, 33] modules, promising results have been achieved.

Recently, GAN prior based methods utilizing pre-trained Generative Adversarial Networks (GANs) have gained attention in the blind face restoration (BFR) or large-scale face super-resolution task, with notable works such as PULSE [23], GPEN [38], GFP-GAN [32], and GLEAN [2]. These methods leverage a pre-trained StyleGAN2 network [18] to generate high-quality facial details and achieve remarkable performance. PULSE [23] adopts a gradient descent algorithm to find the closest latent code of the high-resolution face image to the given low-resolution input image, but it is time-consuming and produces restored images with limited fidelity. Following GAN prior based methods [2, 32, 38], an encoder network with a multi-scale structure is used to map the degraded facial image to a latent code, which is then decoded by StyleGAN to produce a high-quality face image with fast inference time. To preserve the identity of the restored image, these methods often concatenate or fuse the degraded spatial identity features with the

GAN prior features. TFRGAN [37] proposed to leverage the texture information to facilitate the restoration of extremely degraded facial images by using the text and image encoders.

However, the degraded spatial identity features often contain limited identity information of the corresponding high-quality images due to complex degradation. Simply relying on these features that contain substantial degradation but less original identity information can compromise the final reconstruction performance, resulting in undesirable artifacts and insufficient texture details. Moreover, the strong correlation between these degraded spatial identity features has long been overlooked. Furthermore, the fusion of these spatial identity features with decoded GAN priors features is insufficient, resulting in unfaithful facial details and poor fidelity in the restored images.

To address these issues, we propose a method that learns a general representation of degraded spatial identity features by extra memory network to explore the correlation between these features. The memory network is designed to store a comprehensive representation of degraded spatial identity features, including specific facial components such as eyes and noses, as well as auxiliary components such as skin and hair texture at various scales. The memory network aims to enhance the degraded spatial identity features extracted from encoding network, which may contain limited identity information due to complex degradation effects. To achieve high fidelity and realness simultaneously, we propose a coarse-to-fine fusion of the decoded GAN prior features with the memorized spatial features using affine transformation and blending fusion. Our approach aims to fully utilize the information from both degraded spatial identity features and GAN prior features for better restoration performance. The main contributions of our proposed method are summarized as follows.

- We propose a novel approach for blind face restoration that leverages a multi-scale memory network to explore the correlation of degraded spatial identity features extracted from encoding modules. These degraded features can be enhanced with more identity and facial details by adaptively fusing similar facial representations retrieved from the learned memory network. Importantly, our proposed memory network is updated online in an unsupervised manner, obviating the need for pre-training memory network.
- To improve fidelity and realness, we also introduce a coarse-to-fine fusion strategy that merges the decoded GAN prior features with memorized spatial identity features via affine transformation and blending fusion.
- Experimental results on a synthetic dataset and three popular real-world datasets show that our proposed approach achieves comparable or better performance than other state-of-the-art methods for blind face restoration tasks.

2 RELATED WORKS

2.1 Face Restoration.

Since the face images belong to a very small subspace compared to natural images, face restoration solutions can leverage many specific facial priors such as *geometry facial priors* [4, 11, 11] and pre-trained *dictionaries or networks* [19, 32, 34]. Face geometry priors include but are not limited to 2D/3D facial landmark [4, 43], facial parsing maps [3], and facial component heatmaps [39]. Due to the geometry information having to be estimated from degraded face images, the accuracy could be damaged. Besides, geometry only provided position guidance for restoration while no texture information was provided for a realistic reconstruction. A pre-trained facial dictionaries [19, 34, 41, 42] or networks [2, 32, 38] contain abundant details, including facial components and texture. However, the dictionaries or networks are usually pre-trained on HQ face images, and the decoded features from these dictionaries or networks own less fidelity. Differently from motioned methods, we propose a new storage module called memory, which explores the correlation of encoded features and enhances them with less degradation and more details.

2.2 GAN Priors.

Since StyleGAN2 [18] and the previous version [16] have achieved extraordinary generating results with indistinguishable details compared to real images. GAN inversion-based approaches [1, 8, 28, 44] have been proposed to find the most suitable latent codes for given images. More recently, GAN inversion technology has been exploited in face restoration since a pre-trained styleGAN2 has a strong ability for face generation. PULSE [23] proposed to find the latent code of an HQ face image from a given LQ one using a gradient descent algorithm, which is costly and time-consuming. Other GAN prior-based methods [2, 32, 38] use a multi-scale network to encode degraded facial images into latent codes that will be decoded by StyleGAN next. For fidelity, these methods [2, 32, 38] usually simply fused the decoded GAN prior features with the degraded spatial identity features that usually contain a degree of degradation [2, 38]. Deteriorated identity features would damage the final restoration performance, causing unfaithful facial details or poor fidelity. To address this issue, we propose multi-scale memory network to explore the correlation of degraded spatial identity features extracted from encoding modules and enhance these degraded features with more identity via fusing the retrieved similar facial representation from learned memory network.

2.3 Memory Networks.

Memory network is first proposed in [36] for the question-answering system since RNN did not perform as well in long-term memory when dealing with long sequences. More recently, memory networks have improved performance in many computer vision tasks, such as video object segmentation [20, 27], image-to-image translation [15], and image deraining [13]. The use of memory network in computer vision tasks can be roughly divided into two categories: dealing with video sequences [20, 27] and storing regularized information from specific datasets [13, 15]. Jeong *et al* proposed to store class-aware information along with training for image-to-image translation. [13] proposed to adopt memory items to leverage additional real rain data sets unsupervised. Unlike previous work, we propose employing memory network to explore the correlation of degraded spatial identity features encoded from degraded inputs and to enhance these features with more realistic facial details and less degradation. By fully fusing the memorized feature with the decoded GAN prior features, more desirable face reconstruction performance can be obtained.

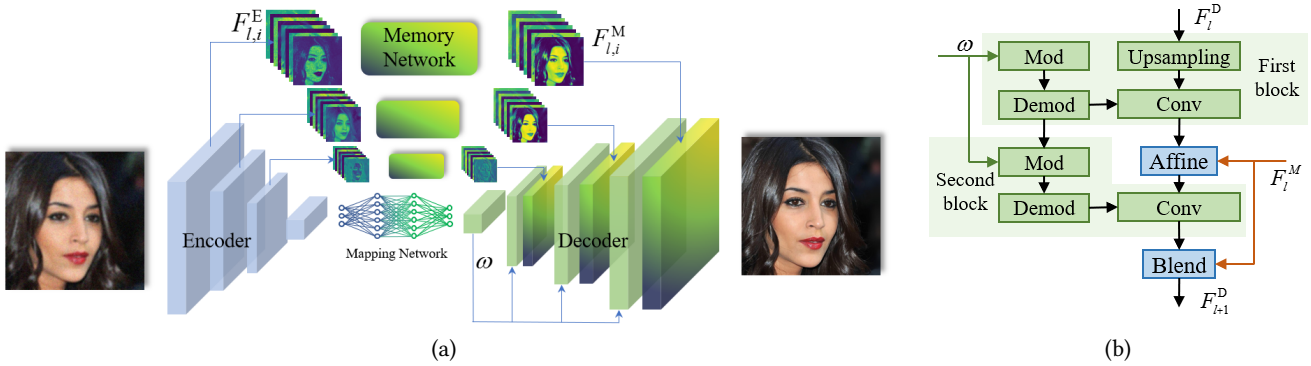


Figure 1: (a)The overall framework of the proposed MemGAN network for real-world blind face restoration; (b)The inner structure of the decoder block of proposed MemGAN in (a).

3 METHODOLOGY

In this section, we first illustrate the overall framework of the proposed method MemGAN. Then, we present how the memory network is updated and similar facial features are retrieved from it. Additionally, we present the proposed coarse-to-fine fusion strategy for integrating the features from memory network and GAN prior in detail. Finally, we describe the training loss functions used in our approach.

3.1 Overall Framework

Our proposed approach, called MemGAN, aims to recover high-quality face images $\hat{\boldsymbol{x}}$ with high fidelity and realism from low-quality face images $\boldsymbol{y}$ with unknown degradation. The proposed framework is illustrated in Fig. 1 (a), and consists of three main components: Encode Blocks, Decoder Blocks, and Memory network. The Encode Blocks are responsible for extracting spatial identity features $\boldsymbol{F}_l^E$ from the degraded images. The encoding features at smallest scale are used to obtain the latent code $\boldsymbol{\omega}$ through an eight-layer fully connected layers (Mapping Network in Fig. 1 (a)). The Decoder Blocks comprise the basic block of StyleGAN2 [18] and proposed fusion modules. StyleGAN2 is a state-of-the-art face generative model that can produce highly realistic facial images that may not exist in the real world with a given latent code. Therefore, we leverage the pre-trained StyleGAN2 as a GAN prior to provide high-quality facial features.

However, due to the difficulty in encoding degraded images into the corresponding HQ latent codes, the recovered HQ images often suffer from fidelity issues. Directly fusing the extracted spatial identity features $\boldsymbol{F}_l^E$ with the StyleGAN characteristics is an intuitive choice [2, 32, 38]. Nevertheless, spatial identity features $\boldsymbol{F}_l^E$ contain a certain degree of degradation and limited identity information, such a direct fusion strategy may generate undesired results with perceived artifacts.

To address this issue, we propose the Memory network and a coarse-to-fine fusion strategy, which exploits the correlation between the degraded spatial identity features and stores the general representation of these features, and enhances them with more identity and facial details via fusing the retrieved similar facial representation from learned memory storage. Specifically, the Memory Network takes the degraded spatial identity features $\boldsymbol{F}_l^E$ as inputs and outputs the memorized spatial identity feature $\boldsymbol{F}_l^M$ with less degradation and more facial details than inputs $\boldsymbol{F}_l^E$, as shown in Fig. 2. Then, the output of memory network $\boldsymbol{F}_l^M$ along with the latent code $\boldsymbol{\omega}$ are fed into decoder blocks. The decoder blocks contain the basic block of pre-trained StyleGAN2 [18] and two fusion modules including affine transformation and blend fusion for better identity reserve. By fully fusing the memorized spatial identity features with the features of GAN prior, the HQ facial images with high fidelity and verisimilitude can be recovered from degraded ones. The retrieving and updating of the memory network will be illustrated in the next section.

3.2 Retrieving and Online Updating of Memory Network

As shown in Fig. 1 (a), a multi-scale memory network is adopted in the MemGAN network. Let superscript l correspond to the scale l in the multi-scale encode-decode structure. The memory module at scale l is denoted as $\boldsymbol{M}_l \in \mathbb{R}^{K,c}$, comprising K memory slots, each with $\boldsymbol{M}_{l,k} \in \mathbb{R}^c$ items, where c is the same as the channel number of the encoding features $\boldsymbol{F}_{l,i}^E \in \mathbb{R}^c$ $(i = 1, \cdots, n)$ and $n = h \times w$. The number of memory slots can be varied as per the requirement. In our implementation, we set $K = h = w$. To enhance the degraded spatial identity features $\boldsymbol{F}_l^E$, we fuse them with similar facial features retrieved from the memory network using soft attention. The memory network is updated in a self-supervised manner by retrieving the most relevant memory network, as we will illustrate next.

Retrieving from memory network. To retrieve similar facial information and facial details stored in the memory network at scale l, we need to calculate the similarity between the degraded spatial identity features $\boldsymbol{F}_{l,i}^E$ and each memory slot $\boldsymbol{M}_{l,k}$. In our implementation, we adopt the cosine distance to measure similarity, which can be formulated as follows:

$$\mathtt{sim}(\boldsymbol{F}_{l,i}^E, \boldsymbol{M}_{l,k}) = \frac{\boldsymbol{F}_{l,i}^E \boldsymbol{M}_{l,k}^T}{||\boldsymbol{F}_{l,i}^E||_2 ||\boldsymbol{M}_{l,k}||_2}. \tag{1}$$

The similarity is then normalized by a softmax function:

$$\boldsymbol{\alpha}_{l,i,k} = \frac{\exp(\mathtt{sim}(\boldsymbol{F}^E_{l,i}, \boldsymbol{M}_{l,k}))}{\sum_{k'}^{K} \exp(\mathtt{sim}(\boldsymbol{F}^E_{l,i}, \boldsymbol{M}_{l,k'}))}, \tag{2}$$

In the end, the spatial identity features are enhanced by a similarity-weighted aggregation of memory network, which can be formulated as:

$$\boldsymbol{F}^M_{l,i} = \sum\nolimits_{k'}^{K} \boldsymbol{\alpha}_{l,i,k'} \boldsymbol{M}_{l,k'}. \tag{3}$$

It is worth mentioning that each slot in the memory network is updated online by spatial features in a self-supervised manner, and no backpropagated gradients are passed into the memory network.

Online updating from spatial identity features. To explore prototypical patterns from abundant degraded spatial identity features, we update the memory network $\boldsymbol{M}$ in a self-supervised manner based on the current degraded spatial identity features $\boldsymbol{F}^E_{l,i}$ and the current memory network. We first calculate the cosine similarity between the spatial identity features $\boldsymbol{F}^E_{l,i}$ and each memory slot $\boldsymbol{M}_{l,k}$ using Eq. (1). Next, we find the most relevant memory item $\boldsymbol{M}_{l,k}$ using

$$\mathtt{index}_{l,i} = \arg\max_{k} \mathtt{sim}(\boldsymbol{F}^E_{l,i}, \boldsymbol{M}_{l,k}). \tag{4}$$

Last, the memory network is updated by the degraded spatial features $\boldsymbol{F}^E_{l,i}$ that have the most relevant item $\mathtt{index}_{l,i} = k$, which can be formulated as

$$\boldsymbol{M}^l_k \leftarrow \lambda \boldsymbol{M}^l_k + (1-\lambda)\frac{\sum_{i=1}^{n} \mathbb{1}(index_{l,i} = k)\boldsymbol{F}^E_{l,i}}{\sum_{i=1}^{n} \mathbb{1}(index_{l,i} = k)}, \tag{5}$$

where $\lambda \in [0, 1]$ and we set $\lambda = 0.999$ in our implementation.

To demonstrate the effectiveness of learning memory network, we have visualized the features before and after the memory network in Fig. 2. The first row shows the features of the largest scale and the second row shows the features of the middle scale. We can observe that the enhanced features shown on the right side of Fig. 2 have less degradation and more facial details, which demonstrates the importance of the memory network modules visually.

Figure 2: Visual comparison of degraded spatial identity features before and after enhancement with the proposed memory network. The left column shows the degraded features before enhancement, while the right column shows the enhanced features. The first row displays the features at the largest scale, and the second row shows the features at the middle scale. Best viewed in color and zoomed in.

3.3 A Coarse-to-fine Fusion Strategy

After enhancing degraded features with more details and texture through the memory network, the next crucial step is to take full advantage of the memorized features. In our approach, we propose a coarse-to-fine fusion strategy, illustrated in Fig.1 (b). This module (referred to as the decoder block in Fig.1 (a)) takes three inputs: the latent code ω, the former layer output $\boldsymbol{F}^D_l$, and the same-scale memorized facial identity features $\boldsymbol{F}^M_l$. The latent code is encoded from degraded images by multiple encode blocks with eight MLP layers.

The features decoded from the first block of the decoder block shown in Fig. 1 (b) are denoted as $\boldsymbol{F}^{GAN1}_l$, whose landmark is far away from the landmark of corresponding HQ face images. To address this, we propose coarsely deforming the features $\boldsymbol{F}^{GAN1}_l$ through an affine transformation based on the memorized spatial identity features. Then, to further enhance the details of the final output of the decoder block $\boldsymbol{F}^D_{l+1}$, we blend the output of the second block in the decoder block, denoted as $\boldsymbol{F}^{GAN2}_l$, with the memorized features using a simple convolution. For brevity, we omit the subscript i of the pixels' position in the following.

Affine Transformation. As shown in Fig. 1 (a), the predicted latent codes ω have been sent to the decoder blocks for decoding. The inner structure of decoder blocks is shown in Fig. 1 (b). However, the features $\boldsymbol{F}^{GAN1}_l$ decoded by StyleGAN2 first block, as shown in Fig. 1 (b), may not be very similar to the original faces. To address this issue, we deform these features $\boldsymbol{F}^{GAN1}_l$ via an affine transformation based on the memorized spatial identity features $\boldsymbol{F}^M_l$. Specifically, we first calculate the scale parameters $\boldsymbol{s}_l$ and bias parameters $\boldsymbol{b}_l$ based on the memorized spatial identity features $\boldsymbol{F}^M_l$, which can be formulated by:

$$\boldsymbol{s}_l, \boldsymbol{b}_l = \mathsf{conv}(\boldsymbol{F}^M_l), \tag{6}$$

where l denotes the scale index. With the calculated scale and bias parameters, we transform the features $\boldsymbol{F}^{GAN1}_l$ encoded by StyleGAN2 first block using the following affine transformation:

$$\boldsymbol{F}^{Affine}_l = \boldsymbol{s}_l \cdot \boldsymbol{F}^{GAN1}_l + \boldsymbol{b}_l, \tag{7}$$

where $\boldsymbol{s}_l$ and $\boldsymbol{b}_l$ have the same size as $\boldsymbol{F}^{GAN1}_l$. This operation scales and shifts the GAN features $\boldsymbol{F}^{GAN1}_l$ based on the memorized identity features $\boldsymbol{F}^M_l$. However, this operation alone does not fully address the identity problem of the recovered facial images since no identity information is involved in the final reconstruction. To overcome this limitation, we propose to blend the features decoded by the StyleGAN2 second block with memorized features that contain rich spatial identity information and facial details.

Blend. The blending fusion is a refinement process aimed at improving fidelity by fusing the memorized features with the features $\boldsymbol{F}^{GAN2}_l$ decoded by StyleGAN2 second block. We use a simple convolution operator to achieve this goal, which can be summarized

as follows:

$$F_{l+1}^{D} = \mathtt{conv}(\mathtt{concat}(F_l^M, F_l^{GAN2})), \tag{8}$$

In this way, the recovered facial images using our coarse-to-fine fusion strategy are more faithful to the original ones.

3.4 Training Loss

Except for the StyleGAN2's basic blocks in decoder blocks is pre-trained, the other components are trained from scratch. Note that the StyleGAN2 in decoder blocks is also trainable. The overall training loss function consists of three parts: the $\mathcal{L}_1$ for maintaining fidelity, perceptual loss $\mathcal{L}_1$ for improving perceived quality, and adversarial loss for restoring realistic textures and realness.

$$\mathcal{L}_1 = ||\hat{\boldsymbol{x}} - \boldsymbol{x}||_1 \tag{9}$$

$$\mathcal{L}_{per} = ||\phi(\hat{\boldsymbol{x}}) - \phi(\boldsymbol{x})||_1, \tag{10}$$

where $\phi(\cdot)$ denotes the feature extractor. In our implementation, we calculated the perceptual loss on VGG19 [29] features of the $\{\mathtt{conv1}, \cdots, \mathtt{conv5}\}$ convolutional layers. Similar to StyleGAN2 [18], logistic loss [7] is adopted:

$$\mathcal{L}_{adv}^{G} = \mathbb{E}_{\hat{\boldsymbol{x}}}[\mathtt{log}(\mathtt{exp}(-D(\hat{\boldsymbol{x}}) + 1)] \tag{11}$$

$$\mathcal{L}_{adv}^{D} = \mathbb{E}_{\boldsymbol{x}}[\mathtt{log}(\mathtt{exp}(D(\hat{\boldsymbol{x}}) + 1) + \mathtt{log}(\mathtt{exp}(-D(\boldsymbol{x}) + 1)] \tag{12}$$

The proposed network is trained by the combination of three losses:

$$\mathcal{L} = \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_{per} + \alpha_3 \mathcal{L}_{adv}, \tag{13}$$

where $\alpha_1 = 0.1, \alpha_2 = 1, \alpha_3 = 0.1$ in this paper.

4 EXPERIMENTS

4.1 Experimental Settings

Training Datasets. The FFHQ dataset [17] containing 70k HQ face images of size 1024×1024, has been used to train BFR approaches [3, 32, 34, 38]. In our implementation, the FFHQ images have been bicubically resized to 512×512 resolution to train our model. The low-quality images are synthesized from high-quality images of FFHQ, and the degradation process can be summarized as:

$$\boldsymbol{y} = \{[(\boldsymbol{x} \circledast \boldsymbol{k}_\delta) \downarrow_s + \boldsymbol{n}_\sigma]_{JPEG_q}\} \uparrow_s \tag{14}$$

First, the HQ image $\boldsymbol{x}$ is convolved with the Gaussian blur kernel $\boldsymbol{k}_\delta$, and δ denotes the parameter related to the degree of blur. Then, a bilinear downsampling operator is adopted with s scale. The JPEG compression with quality q is followed after adding the additive white Gaussian noise with variance σ. Finally, to maintain the consistent spatial resolution of the BFR, the degraded images will be bilinearly upsampled to original size. Following [32, 34], the degradation parameters δ, s, σ, and q are randomly selected from $\{0.2 : 10\}$, $\{1 : 8\}$, $\{0 : 15\}$, $\{60 : 100\}$, respectively. Even though our model is trained on the synthetic date, it has a generalization ability to degrade face images in the real world during testing.

Testing Datasets. Following [32, 34], a synthetic data set called **CelebA-Test** and three real-world data sets: **LFW**, **CelebChild**, and **WebPhoto**, have been used to evaluate our approach. CelebA-Test contains 2824 images from the CelebA-HQ [22] testing partition. The degraded images are synthesized by the same degradation process of training. LFW [12] contains 15,154 low quality images from the wild that are all used for our evaluation. CelebChild [32] is consist of 180 child faces of celebrities and WebPhoto [32] consists of 188 low-quality photos in real life, which are collected by [32] from Internet.

Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓
Input	99.09	**24.61**	0.6180	0.6276
DFDNet [19]	42.39	23.23	0.6351	0.4891
PULSE [23]	43.58	21.48	0.5791	0.4893
wan *et al* [31]	77.45	24.03	0.6065	0.5379
PSFRGAN[3]	21.03	24.19	0.6069	0.3960
GPEN [38]	<u>14.41</u>	24.14	0.6104	0.3593
GFP-GAN [32]	16.42	24.17	**0.6403**	**0.3405**
CodeFormer [42]	16.97	23.39	0.6110	0.3503
DiFace [40]	15.50	23.84	0.6275	0.3825
DR2 [35]	31.15	23.96	<u>0.6352</u>	0.4182
MemGAN(ours)	**13.01**	<u>24.28</u>	0.6249	<u>0.3487</u>
GT	2.27	∞	1	0

Table 1: Average FID, PSNR, SSIM, and LPIPS results on CelebA-Test dataset [22]. The best results are shown in bold and the second-best results are shown in <u>underline</u>.

Evaluation Metrics. For the dataset CelebA-Test that provides LQ-HQ pairs for evaluation, we employ PSNR, SSIM, and LPIPS metrics to measure the restoration performance. For the datasets without ground truth images, we adopt widely-used FID [10] metrics to measure the statistical distance between the restoration results and an HQ reference face dataset. We choose the ground truth images of CelebA-Test as the HQ reference face dataset. Besides, the FID score of CelebA-Test is also provided for comparison.

Implementation Details. The pre-trained StyleGAN2 [18] with 512×512 is used as a pre-trained weight. We randomly select 12 RGB LR images sized by 512×512 as a batch input. The learning rate is set as 0.002. We train the model for 300k iterations. Our network is implemented under the Pytorch framework and the training time takes less than 2 days using 4 NVIDIA 3090Ti GPUs.

4.2 Comparsion with SOTA Methods

We have compared our proposed approach MemGAN with several state-of-the-art real-world blind face restoration methods, such as DFDNet [19], PULSE [23], wan *et al* [31], PSFRGAN[3], GPEN [38], GFP-GAN [32], VQFR [9], CodeFormer [42], DiFace [40] and DR2 [35]. The comparison results mainly consist of two parts: the synthetic dataset: CelebA-Test [22] and three Real-World Dataset: LFW [12], CelebChild [32], and WebPhoto [32].

Synthetic Dataset: CelebA-Test. The quantitative comparison on CelebA-Test is shown in Tab. 1, from which we can see that our proposed approach has achieved the best performance compared to other state-of-art methods in terms of FID and PSNR. It indicates the distribution of our restored facial images is the closest to the real ones and our restored facial images are the most similar to ground truth facial images simultaneously. In terms of the perceived

Figure 3: Visual quality comparison of CelebA-Test. Zoom in for better view.

quality assessment metric LPIPS, our proposed approach also has comparable results.

Visual qualitative comparisons are shown in Fig. 3. The CelebA-Test provides ground truth for reference, making it easier to compare the performance of different SOTA methods. DFDNet [19] restores faces with perceived degradation, resulting in low-quality results. The facial images restored by PULSE [23] are visually pleasant as shown in Fig. 3, but these images suffer from severe fidelity problem which is not similar to the original ones. PSFRGAN [3] generates faces with obvious and unpleasant artifacts. The StyleGAN-based method, GFPGAN has trouble restoring faithful facial images to the original ones such as the color and shape of the pupil as shown in Fig. 3. This may be caused by fusing degraded spatial features with decoded StyleGAN features directly. Different from GFPGAN, our approach enhances the degraded identity with less degradation and more facial details via retrieving similar features from memory network, achieving the best visual results with both realness and identity. Taking the third row of Fig. 3 as an example, our method recovers the hair more visually pleasant than GFPGAN. In contrast, GFPGAN generates noticeable artifacts in the hair region, which negatively impacts the visual quality of the final output.

Real-World Datasets: LFW, CelebChild, and Web Photo. The quantitative comparison results of three real-world datasets are shown in Tab. 3. Since no ground truth faces of three real-world datasets are available, we adopt FID as our metric. The calculation of FID requires a high-quality facial dataset for reference, which does not have to be paired with degraded facial images. In our calculation, the ground truth images of CelebA-Test have been used as a reference dataset. As shown in Tab. 3, our method achieved the best performance in real-world huge testing set LFW and smaller testing set WebPhoto, showing that the distribution of restored facial images by our proposed method is closest to the real high-quality ones. In the dataset CelebChild, we are only inferior to the VQFR [9].

The visual comparisons are shown in Fig. 4. No ground truth images of these three real-world datasets are available. Therefore, the assessment in terms of identity of restored facial images relies on the degraded inputs. In Fig. 4, the first four rows show the comparisons from LFW [12]; the fifth rows show the comparisons from CelebChild [32]; the sixth rows show the comparisons from WebPhoto [32]. In particular, we show some zoomed-in images of important components of the face images, such as mouth, hair, eyes, and *et al.* As shown in the second row of Fig. 4, our method recovers a very realistic hair while other competing methods fail with unpleasant artifacts or blurring results, demonstrating the superiority of our proposed method that achieves realness and identity simultaneously. A similar phenomenon can also be found in CelebChild dataset shown in the fifth row in the Fig. 4. Taking the mouth as an example, the first, third, and fifth rows show the visual comparison around the mouth. Compared with other methods, our approach achieves the results of the most reasonable and fewest artifacts.

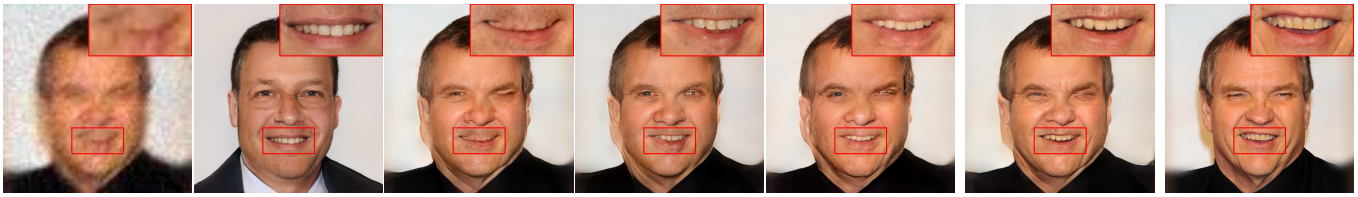
4.3 Ablation Study

We have conducted some ablation study experiments to verify the effectiveness of the proposed modules: Memory Network, Affine Transformation, and Blend Fusion. The comparison of ablation study experiments has been evaluated in the synthetic dataset: CelebA-Test [22], which provides high-quality images for reference. The quantitative results of the ablation study in terms of FID/PSNR/SSIM/LPIPS are shown in Tab. 2, and the visual comparisons are shown in Fig. 5. The investigation can be divided into three parts:

Figure 4: Visual quality comparison of three real-world datasets. The first four rows show the comparisons from LFW [12]; the fifth row shows the comparisons from CelebChild [32]; the sixth row shows the comparisons from WebPhoto [32]. Zoom in for better view.

Memory Network. This subsection analyzes the effect of the Memory Network. The comparison experiments are shown in Case 2 (without Memory Network) and Case 5 (with Memory Network) in Tab. 2. It can be observed that the Memory Network boosts the reconstruction performance in all comparing metrics. The improvement in FID and LPIPS metrics indicates that recovered images with Memory Network are more similar to the real ones than those without Memory Network. The visual comparison in Fig. 5 (c) and (f) shows that in Case 2 without Memory Network, the teeth are not restored at all. With Memory Network, the teeth are realistically recovered. The Memory Network also helps general representations of degraded spatial identity features to contribute to the final reconstruction. The visual comparison of features before and after Memory Network is shown in Fig. 2. It can be observed that cleaner features, shown on the right of Fig. 2, with less degradation and more facial details can be obtained after being enhanced by the Memory Network modules.

Affine Transformation. As shown in Fig. 1 (b), the affine transformation operation is used as the first fusion between memorized



(a) LQ (b) Case1 (c) Case2 (d) Case3 (e) Case4 (f) Case5 (MemGAN) (g) GT

Figure 5: Visual quality comparison of ablation study. Zoom in for better view. The Case $*$ responds to the situation in Tab. 2.

	Memory Network	Affine Transformation	Blend Fusion	FID ↓	PSNR↑	SSIM↑	LPIPS↓
Case1				34.94	15.25	0.4456	0.5210
Case2		✓	✓	17.22	23.95	0.6154	0.3628
Case3	✓		✓	16.34	24.04	0.6162	0.3626
Case4	✓	✓		15.56	23.00	0.6101	0.3739
Case5 (MemGAN)	✓	✓	✓	**13.46**	**24.28**	**0.6249**	**0.3487**

Table 2: The ablation study of proposed modules. The average FID, PSNR, SSIM, and LPIPS results are evaluated on CelebA-Test dataset. The best results are shown in bold.

Methods	**LFW**	**CelebChild**	**WebPhoto**
Input	122.58	136.44	176.53
DFDNet [19]	66.68	115.20	124.13
wan *et al* [31]	109.37	166.44	159.64
PSFRGAN [3]	67.08	136.46	<u>111.87</u>
GFP-GAN [32]	61.91	128.51	125.69
VQFR [9]	<u>57.33</u>	**110.67**	113.32
CodeFormer [42]	58.85	122.11	114.67
DifFace [40]	58.94	121.29	125.92
DR2 [35]	59.70	129.50	140.01
MemGAN(ours)	**56.60**	<u>115.79</u>	**106.83**

Table 3: Average FID (the lower, the better) on three real-world datasets. The best results are shown in bold and the second-best results are shown in <u>underline</u>.

features and decoded StyleGAN2 features, aiming to affine transform the StyleGAN2 features based on memorized spatial identity features. When comparing Case 3 and Case 5 in Tab. 2, the affine transformation can improve all comparing metrics. The visual comparison is shown in Fig. 5 (d) and (f). With the affine transformation, we can see that Case 5 (MemGAN) can recover a more correct spatial position, such as mouth, teeth, and face contour. It can demonstrate that the affine transformation fusion strategy can avoid or relieve the distortion of important parts of the faces.

Blend Fusion. The Blend Fusion module is proposed to improve the identity and facial details of the final construction. When comparing Case 4 and Case 5 in Tab. 2, adding a Blend Fusion operation can improve the final recovered results in terms of all comparing metrics. The visual comparison is shown in Fig. 5 (e) and (f). With the second Blend Fusion strategy, the recovered face is more faithful to the original face, with more details.

5 CONCLUSION

In this paper, we propose a novel approach called MemGAN, which utilizes a multi-scale memory network to explore the correlation of degraded spatial identity features and enhance their identity via fusion with retrieved similar facial representations from the learned memory network. Our proposed coarse-to-fine fusion strategy integrates the decoded GAN prior features and memorized spatial identity features using affine transformation and blending fusion to achieve both fidelity and realness. Additionally, our memory network is updated unsupervised online along with other modules of our approach. Our method demonstrates the effectiveness of exploiting the correlation of degraded spatial identity features and provides a promising solution for blind face restoration.

ACKNOWLEDGEMENT

This work was supported in part by the Natural Science Foundation of China under Grant 61991451 and Grant 61836008. Xin Li's work is partially supported by the NSF under grant IIS-2114664, CMMI-2146076, and the WV Higher Education Policy Commission Grant (HEPC.dsr.23.7).

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.
[2] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. 2021. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14245–14254.
[3] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. 2021. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11896–11905.
[4] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2492–2501.
[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2015. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE*

international conference on computer vision. 576–584.

[6] Weisheng Dong, Xin Li, Lei Zhang, and Guangming Shi. 2011. Sparsity-based image denoising via dictionary learning and structural clustering. In *CVPR 2011*. IEEE, 457–464.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[8] Jinjin Gu, Yujun Shen, and Bolei Zhou. 2020. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3012–3021.

[9] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. 2022. VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In *ECCV*.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[11] Xiaobin Hu, Wenqi Ren, Jiaolong Yang, Xiaochun Cao, David P Wipf, Bjoern Menze, Xin Tong, and Hongbin Zha. 2021. Face Restoration via Plug-and-Play 3D Facial Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[12] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. 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*.

[13] Huaibo Huang, Aijing Yu, and Ran He. 2021. Memory oriented transfer learning for semi-supervised image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7732–7741.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).

[15] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. 2021. Memory-guided unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6558–6567.

[16] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[17] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[19] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. 2020. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*. Springer, 399–415.

[20] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. 2021. Video object segmentation with dynamic memory networks and adaptive object alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8065–8074.

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 136–144.

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[23] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2437–2445.

[24] Qian Ning, Weisheng Dong, Xin Li, and Jinjian Wu. 2022. Searching Efficient Model-Guided Deep Network for Image Denoising. *IEEE Transactions on Image Processing* 32 (2022), 668–681.

[25] Qian Ning, Weisheng Dong, Guangming Shi, Leida Li, and Xin Li. 2020. Accurate and lightweight image super-resolution with model-guided deep unfolding network. *IEEE Journal of Selected Topics in Signal Processing* 15, 2 (2020), 240–252.

[26] Qian Ning, Jingzhu Tang, Fangfang Wu, Weisheng Dong, Xin Li, and Guangming Shi. 2022. Learning degradation uncertainty for unsupervised real-world image super-resolution. In *Proc. 31st Int. Joint Conferences Artif. Intell.* 1261–1267.

[27] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9226–9235.

[28] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. 2021. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[29] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[30] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. 2018. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8174–8182.

[31] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. 2020. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2747–2757.

[32] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9168–9178.

[33] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 606–615.

[34] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. 2022. RestoreFormer: High-Quality Blind Face Restoration From Undegraded Key-Value Pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17512–17521.

[35] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. 2023. DR2: Diffusion-based Robust Degradation Remover for Blind Face Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1704–1713.

[36] Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* (2014).

[37] Chengxing Xie, Qian Ning, Weisheng Dong, and Guangming Shi. 2023. TFRGAN: Leveraging Text Information for Blind Face Restoration With Extreme Degradation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2534–2544.

[38] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. 2021. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 672–681.

[39] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. 2018. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*. 217–233.

[40] Zongsheng Yue and Chen Change Loy. 2022. DifFace: Blind Face Restoration with Diffused Error Contraction. *arXiv preprint arXiv:2212.06512* (2022).

[41] Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. 2022. Rethinking Deep Face Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7652–7661.

[42] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. 2022. Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. In *NeurIPS*.

[43] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. 2022. Blind Face Restoration via Integrating Face Shape and Generative Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7662–7671.

[44] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain gan inversion for real image editing. In *European conference on computer vision*. Springer, 592–608.