ShowFace: Coordinated Face Inpainting with Memory-Disentangled Refinement Networks

Zhuojie Wu[1]
zhuojiewu@bupt.edu.cn
Xingqun Qi[1]
xingqunqi@gmail.com
Zijian Wang[1]
wangzijianbupt@bupt.edu.cn
Wanting Zhou[1]
wanting.zhou@bupt.edu.cn
Kun Yuan[2]
yuankun03@kuaishou.com
Muyi Sun[3, ✉]
muyi.sun@cripac.ia.ac.cn
Zhenan Sun[3]
znsun@nlpr.ia.ac.cn

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), Beijing, China

[2] Kuaishou Technology, Shenzhen, China

[3] Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China

Abstract

Face inpainting aims to complete the corrupted regions of the face images, which requires coordination between the completed areas and the non-corrupted areas. Recently, memory-oriented methods illustrate great prospects in the generation related tasks by introducing an external memory module to improve image coordination. However, such methods still have limitations in restoring the consistency and continuity for specific facial semantic parts. In this paper, we propose the coarse-to-fine Memory-Disentangled Refinement Networks (MDRNets) for coordinated face inpainting, in which two collaborative modules are integrated, Disentangled Memory Module (DMM) and Mask-Region Enhanced Module (MREM). Specifically, the DMM establishes **a group of disentangled memory blocks to store the semantic-decoupled face representations**, which could provide the most relevant information to refine the semantic-level coordination. The MREM involves a **masked correlation mining mechanism** to enhance the feature relationships into the corrupted regions, which could also make up for the correlation loss caused by memory disentanglement. Furthermore, to better improve the inter-coordination between the corrupted and non-corrupted regions and enhance the intra-coordination in corrupted regions, we design **InCo^2 Loss, a pair of similarity based losses** to constrain the feature consistency. Eventually, extensive experiments conducted on CelebA-HQ and FFHQ datasets demonstrate the superiority of our MDRNets compared with previous State-Of-The-Art methods.

✉ Muyi Sun is the corresponding author.

1 Introduction

Face inpainting is an ill-posed problem, which aims to restore the corrupted regions with coordinated contents as long as they appear plausible[13, 14, 27, 37, 40]. Since the corrupted regions traverse multiple semantic parts, the coordinated inpainting generally requires consistency within each semantic and coordination between different semantic parts. Recently, face inpainting has shown great potential in real-world applications such as interactive face editing and occluded face recognition. However, it is still challenging for coordinated face recovery when the missing regions are large or the face contents are complex. Figure 1 shows this inpainting task and illustrates several results of our method, which achieve coordinated face inpainting with large masked regions.

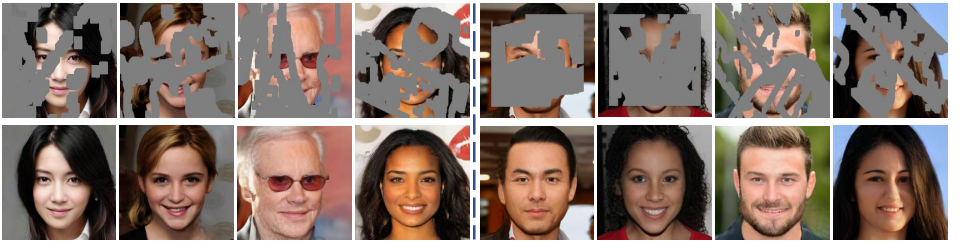

Figure 1: Randomly sampled results from our MDRNets. The first four columns are from the CelebA-HQ[11] dataset, and the last four columns are from FFHQ[12] dataset.

From the perspective of technology, face inpainting could be roughly divided into two categories in recent years: patch-based methods [1] and deep learning based methods [14]. Specifically, the patch-based methods [25, 32, 43] generally sample image patches from the remaining image regions and fuse these patches to recover the missing areas, which could synthesize highly-textured complement contents. However, the above methods are difficult to generate semantically reasonable results due to the lack of high-level image understanding. Meanwhile, the patch matching process is usually time consuming and laborious. Furthermore, the large corrupted regions with insufficient remaining patches lead to over-smooth completed results.

Recently, great progress has been made in inpainting tasks with the remarkable development of Generative Adversarial Networks(GANs) [8]. Meanwhile, numerous researchers employ GANs in face inpainting[14, 17, 30, 33, 36]. For the specificity of the inpainting task, the vanilla convolution is upgraded in [15, 30, 37], which change the conventional convolution mechanism and pay more attention to valid pixels. Additionally, many studies improve the network to better recover the global structure by introducing relevant structural priors [4, 20, 22, 26, 38]. Meanwhile, some researchers [16, 31, 36] combine the advantages of the patch-based methods and deep learning based methods, which could deliver the inpainting contents with both detailed textures and plausible semantics. However, the above methods ignore the consistency and continuity for specific facial semantic parts, which is an important problem for coordinated face inpainting.

Inspired by the above studies, we propose the coarse-to-fine Memory-Disentangled Refinement Networks (MDRNets) for coordinated face inpainting. Firstly, the coarse network generates a coarse face, which could produce reasonable structural priors. Then, we design a **Disentangled Memory Module (DMM)** to store the semantic-aware decoupled face latent vectors from the non-corrupted regions. With this design, the disentangled semantic-level

latent vectors ensure the coordination within each semantic. Next, we propose a **Mask-Region Enhanced Module (MREM)** to enhance the feature relationships into the corrupted regions. The MREM involves a masked correlation mining mechanism to compute the relationships between the completed and the non-corrupted regions. Finally, we utilize the output of MREM to refine the coarse face in the guided refinement network. For improving the intra-coordination in corrupted regions and inter-coordination between corrupted and non-corrupted regions, we design **InCo^2 Loss, a pair of similarity based losses** to constrain the feature consistency. Specifically, we construct two types of similarity matrices to mine deeper feature correlations from the corrupted and non-corrupted regions.

The main contributions can be summarized as follows:

- We propose MDRNets for coordinated face inpainting.
- We design two collaborative modules, DMM and MREM, which achieve the memory disentanglement for semantic-level coordination and enhance the feature relationships for face inpainting.
- We propose InCo^2 Loss, a pair of similarity based losses to further improve the intra-coordination in corrupted regions and the inter-coordination between the corrupted and non-corrupted regions.
- Both the qualitative and quantitative results on CelebA-HQ and FFHQ datasets demonstrate our method effectiveness which achieves State-Of-The-Art performance.

2 Related Work

2.1 Face Inpainting

Face inpainting has made tremendous progress in the past few years. In previous patch-based methods, Zhuang *et al.* [43] and Tang *et al.* [25] extract prototypical image patches to fill the missing areas. Xu *et al.*[31] utilize the similarity matrix to seek patches for consistent texture generation. However, the patch-based methods are difficult to find suitable contents when the corrupted regions is foreground and large. Then, great efforts are made in early deep learning methods to maintain the image consistency and restore irregular masks. Yu *et al.* [36] establish the contextual relationship into the face inpainting networks. Liu *et al.* [15] propose partial convolution for irregular mask to filter out invalid pixels. Yu *et al.* [37] design a learnable dynamic feature selection mechanism, which generalizes the partial convolution. However, these early deep learning methods are limited in maintaining global consistency among face components, and the completed areas are generally blurry. Recently, some methods are designed to integrate the face priors or new network architectures. Li *et al.* [13] propose SymmFCNet, which use the symmetry of face to recover facial details. Liu *et al.* [18] introduce probabilistic diversity map, which controls the diversity extent of the completed faces. Peng *et al.* [22] and Guo *et al.* [4] utilize facial structure and texture constraints to guide the inpainting network. Wan *et al.* [26] and Yu *et al.* [38] employ autoregressive transformers to inpaint diverse faces.

2.2 Memory Networks

Extensive deep learning methods possess the ability of memory, such as RNN [19], LSTM[6] and GRU[2]. However, they are all limited in the long-term memory of information. To overcome this shortcoming, Weston *et al.* [29] first propose memory networks, which employ

explicit storage and attention mechanism to model the long-term information more effectively. And due to the high efficiency for feature storage, memory networks have become popular in the field of image generation. Yoo *et al.* [34] present a memory-augmented colorization network to produce high-quality image colorization with limited data. Huang *et al.* [7] employ a dynamic memory block to record the prototypical patterns of rain degradations for rain removal. Zhu *et al.* [41] introduce a multimodal memory module to refine blurred images for text-to-image generation. Qi *et al.* [23] design a latent memory unit to preserve the core storyline and history information for visual storytelling. In the face inpainting task, Xu *et al.* [31] firstly propose a patch-based memory to enhance the completed image texture.

3 METHOD

In this section, we present our method in detail. Firstly, we introduce the overall MDRNets. Then, we give the details of the specific components in the networks, especially the DMM and MREM. Finally, the total objective functions of this model are described.

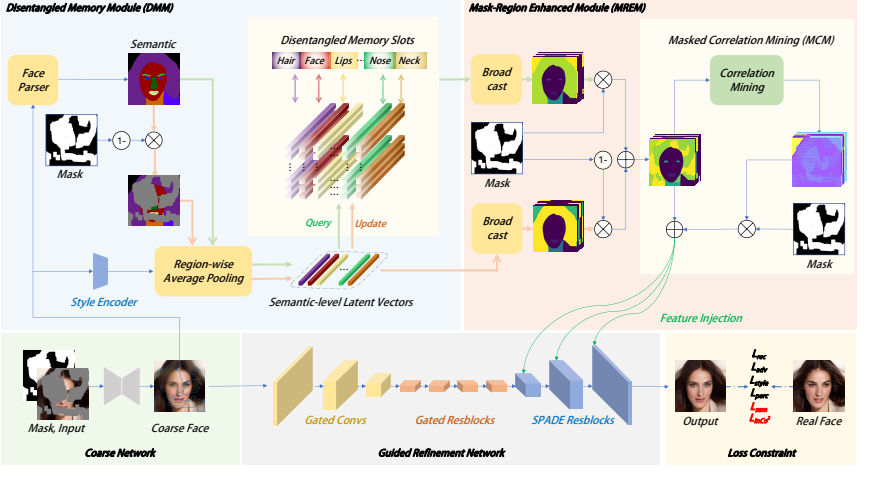

Figure 2: An overview of MDRNets. ⊖, ⊕ and ⊗ denote the operations of 1-Mask, element-wise addition and element-wise multiplication respectively.

3.1 Overview

The proposed MDRNets is shown in Figure 2. Firstly, given a masked image I and the corresponding mask M, we employ the pre-trained partial convolution based coarse network to generate the coarse face P. Then we leverage the face parser [35] to obtain the corresponding semantic map S of the coarse result. To finely recover each semantic part of the face and maintain semantic coordination, the DMM (i.e. the memory $\mathbb{M}$) is proposed to store the semantic-aware latent vectors V from the non-corrupted regions. Then, the most relevant memory slots in $\mathbb{M}$ could be retrieved using the semantic-level latent vectors Q of the coarse face as queries. To enhance the feature, the MREM is proposed to construct a correlation map, which could fuse the features of the non-corrupted regions into the corrupted regions. Eventually, the generated features after MREM are injected into the Guided Refinement Network through SPADE [21] to get the coordinated face $\hat{I}$.

3.2 Disentangled Memory Module

To generate semantic-level coordinated faces, we propose the DMM to establish a group of disentangled memory blocks, which stores the semantic-decoupled face representations. As illustrated in Figure 2, we employ the Style Encoder to extract the style feature maps $F_s \in \mathbb{R}^{c\times h\times w}$ from the coarse face P. Meanwhile, the face parser is used to obtain the corresponding semantic map S from P, which contains 14 different semantic categories in face images (*e.g., skin, eye*). Then, we employ region-wise average pooling [42] to obtain latent vectors $Q \in \mathbb{R}^{n\times c}$ and $V \in \mathbb{R}^{n\times c}$, where n represents the number of semantic categories.

Memory blocks. We establish the disentangled memory blocks (14 blocks for 14 semantic categories) to store the semantic-aware latent vectors V, which represents the non-corrupted and accurate latent representations of the facial parts. Specifically, the proposed memory $\mathbb{M} \in \mathbb{R}^{n\times m\times c}$ consists of $n = 14$ memory blocks, in which each memory block contains m memory slots $e_{ij} \in \mathbb{R}^c$. Taking each semantic-level latent vector in Q as a query, we could retrieve a relevant representation from its corresponding memory block. This memory-based representation is obtained by integrating the m semantic-related memory slots with soft scores. Meanwhile, we could update the semantic-level $\mathbb{M}$ by the semantic-aware V.

Memory Updating. The update of memory $\mathbb{M}$ is based on the similarity between the latent vectors in Q and the corresponding memory slots. To begin with, we compute the i-th semantic cosine similarity γ_{ij} between Q_i and the j-th memory slot e_{ij}, defined as

$$\gamma_{ij} = \frac{e_{ij}Q_i^T}{\|e_{ij}\| \, \|Q_i\|} \tag{1}$$

Then, we retrieve the memory slot $e_{i\phi_j}$, which is most relevant with Q_i in each training batch.

$$k_j = \underset{i}{argmax}(\gamma_{ij}) \tag{2}$$

To ensure the authenticity of the memory, we use the **non-corrupted latent vector** V_i from the non-corrupted regions in each training batch to update the memory slot e_{ik_j}, **which is also most similar and shares the same semantic with the query** Q_i.

$$e_{ik_j} \leftarrow \alpha e_{ik_j} + (1-\alpha)V_i \tag{3}$$

where $\alpha \in [0,1]$ is a decay rate.

Memory Reading. After updating the memory $\mathbb{M}$, we reconstruct memory-based latent vectors $\hat{Q}_i$, which is most relevant to Q_i. What's more, we employ soft scores to aggregate memory slots for end-to-end training. To begin with, the cosine similarity matrix $\Upsilon = \{\gamma_{ij} | i = 1,...,n, j = 1,...,m\}$ is computed by Eq.1 again. Then, the soft scores $A = \{a_{ij} | i = 1,...,n, j = 1,...,m\}$ are formulated by a softmax operation.

$$a_{ij} = \frac{exp(\gamma_{ij})}{\sum_{j=1}^{m} exp(\gamma_{ij})} \tag{4}$$

Finally, the memory-based latent vectors $\hat{Q}_i$ is constructed by aggregating memory slots with the soft scores.

$$\hat{Q}_i = \sum_{j}^{m} a_{ij} e_{ij} \tag{5}$$

3.3 Mask-Region Enhanced Module

To enhance the feature representation of the corrupted regions, we propose MREM, which consists of feature fusion and Masked Correlation Mining (MCM). To begin with, we broadcast memory-based latent vectors $\hat{Q}$ to semantic map S, which obtain memory-based feature maps. Meanwhile, we obtain F_V by broadcasting V to semantic map S. Then, we employ the mask to achieve feature fusion, which ensures the fused features both come from the "real" image features and share great similarity at the semantic level. The above processes are shown in Figure 2.

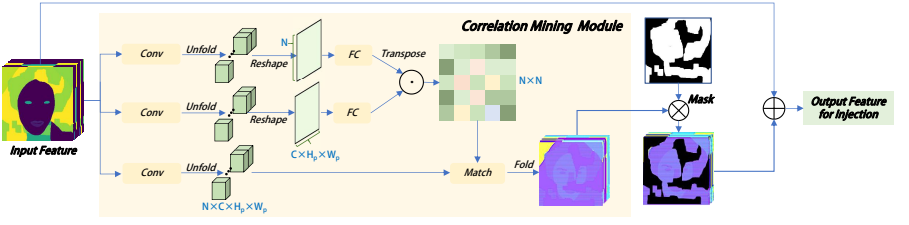

Figure 3: Detailed illustration of the Masked Correlation Mining (MCM). $\odot$, $\oplus$ and $\otimes$ denote the dot product, element-wise addition and element-wise multiplication respectively.

To focus on feature relationships, we further design the MCM, which consists of a correlation mining module and mask multiplication. The correlation mining module contains three branches, in which the first two branches compute the correlations within the features and then match the third branch. To begin with, We apply the 1×1 convolutional layer to transform the input features into two independent representations, and then utilize the unfold operation to extract N feature patches $\mathcal{P} \in \mathbb{R}^{C \times H_p \times W_p}$. Next, each feature patch is reshaped into a feature vector. The similarity matrix $\Phi \in \mathbb{R}^{N \times N}$ representing the correlations between each patch can be computed by dot product. Thus, we could update each patch by the similarity matrix Φ. Then, the mask multiplication preserves the feature enhancement of the corrupted regions. Finally, the correlation-enhanced corrupted regions are fused with the input features by element-wise addition.

3.4 Guided Refinement Network

The Guided Refinement Network consists of gated convolutional layers [37], gated ResBlocks, and SPADE ResBlocks [21]. Firstly, we encode the P to provide the texture of the non-corrupted regions for the final face generation. Then, the fused features F_f after MREM are injected into the Guided Refinement Network by SPADE as shown in Figure 2, which facilitate the final coordinated face. The more detailed description of the Guided Refinement Network is given in supplementary materials.

3.5 Objective Functions

In this section, we introduce the proposed InCo^2 loss in detail and present the objective functions of our method.

In face inpainting, it is reasonable to focus on corrupted region reconstruction. Meanwhile, coordinated face inpainting between the completed and the non-corrupted areas is also a key point. In this paper, we propose InCo^2 Loss to further constrain the feature consistency for face coordination. Specifically, the **InCo^2 Loss** includes a pair of similarity based losses, **In**tra-class **Co**ordination loss and **In**ter-class **Co**ordination loss. The intra-class coordination

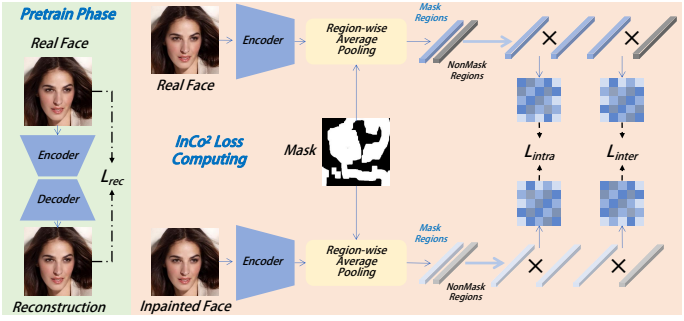

Figure 4: Detailed illustration of InCo^2 loss.

requires the coordinated relationships among the various semantic features in corrupted regions. Similarly, inter-class coordination requires coordinated feature relationships between the completed regions and the non-corrupted regions.

Concretely, there are two steps to establish the InCo^2 Loss, as shown in Figure 4. In the pretrain phase, we employ an encoder-decoder based reconstruction network to obtain the implicit representations of face from the middle layer. Then we employ the pretrained encoder and the mask to obtain two representations $\mathcal{M}(\cdot)$ and $\hat{\mathcal{M}}(\cdot)$ of the mask and non-mask regions by region-wise average pooling[42], respectively. The intra-class coordination loss and the inter-class coordination loss are defined as:

$$\mathcal{L}_{intra} = \left\| \mathcal{M}(\hat{I}) \times \mathcal{M}(\hat{I})^T - \mathcal{M}(I_{gt}) \times \mathcal{M}(I_{gt})^T \right\|_1 \tag{6}$$

$$\mathcal{L}_{inter} = \left\| \mathcal{M}(\hat{I}) \times \hat{\mathcal{M}}(\hat{I})^T - \mathcal{M}(I_{gt}) \times \hat{\mathcal{M}}(I_{gt})^T \right\|_1 \tag{7}$$

Therefore, InCo^2 Loss is defined as:

$$\mathcal{L}_{InCo^2} = \mathcal{L}_{intra} + \mathcal{L}_{inter} \tag{8}$$

In addition to the InCo^2 Loss, we follow the previous work [22, 26] and utilize the semantic loss[14], reconstruction loss[18], perceptual loss[10], style loss[8], adversarial loss[9] and total variation loss[15] to optimize our network. The more detailed description of the objective function is given in supplementary materials.

4 Experimental Settings

Datasets and Evaluation Metrics. We evaluate the proposed method on CelebA-HQ [11] and FFHQ [12]. We follow the split in [37] to produce 28,000 training images and 2,000 validation images in CelebA-HQ. For FFHQ, we preserve the last 2,000 images for test, and use the rest images for train. Irregular masks provided by [15] are employed for both training and evaluation. The $L1$ error, Fréchet Inception Distance (FID) [5], Peak Signal-to-Noise Ratio (PSNR), Structure Similarity (SSIM) [28] are used to evaluate the quality of the results. The $L1$ error, PSNR and SSIM compare the differences between the completed image and ground truth. The FID calculates the distance of feature distributions between the completed face and ground truth.

Implementation Details. The proposed method is implemented in PyTorch with 4 Nvidia Titan Xp GPUs. The image and mask are resized to 256 $\times$ 256 for training and evaluation. Our model is optimized using Adam optimizer with β_1=0.9 and β_2=0.99. We

train the model for 45 epochs with the batchsize of 8. The learning rate is set to 2e-4. For coarse network, We replace the vanilla convolution of the U-Net architecture [24] with partial convolution [15] as the coarse network. The coarse network is trained on CelebA-HQ for 100 epochs, and other settings are the same as the MDRNets. For the reconstruction network in Figure 4, we train the network for 30 epochs, with the same settings as above.

4.1 Qualitative Analysis

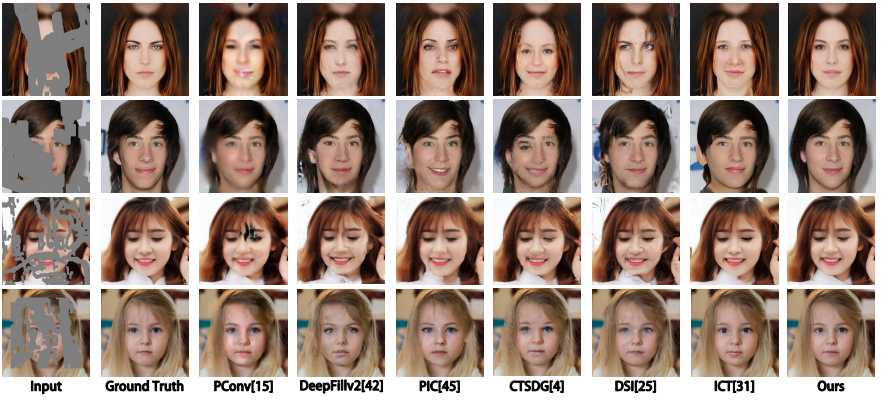

Figure 5: Randomly sampled results of our MDRNets compared with the previous SOTA methods. Zoom in for better details.

We compare our methods with previous state-of-the-art approaches, including PConv [15], DeepFillv2 [37], PIC [39], CTSDG [4], DSI [22] and ICT [26]. All the results are obtained by using pre-trained models or implementation code published by the authors. We show the results of qualitative comparisons in Figure 5. PConv and DeepFillv2 generate blurry results since these models can not capture valid contextual information. PIC generates reasonable facial structures. However, the results of PIC suffer from artifacts due to the lack of adequate correlation. CTSDG and DSI obtain distorted faces since these models use low-level structural information, which is incomplete in wide corrupted regions. ICT could generate relatively satisfactory results. However, the results of ICT still have limitations in detailed textures since the model cannot perform fine restoration of each semantics. Compared with these methods, our model achieves better results on both detailed textures and face coordination. More qualitative results are presented in the supplementary materials.

4.2 Quantitative Analysis

As shown in Table 1, we quantitatively evaluate the proposed method at irregular mask ratios of 1-20%, 20-40% and 40-60%. As we can see, our proposed method outperforms other State-Of-The-Art methods on CelebA-HQ and FFHQ datasets. Especially, under the largest mask ratio, our method has **distinct improvements** compared with other methods. Specifically, the $L1$ error and FID are reduced by 0.963% and 4.661. Meanwhile, the PSNR and SSIM are improved by 1.687 and 0.035, compared to the sub-optimal result on the CelebA-HQ. Similarly, the $L1$ error and FID are reduced by 0.501% and 0.401. The PSNR and SSIM are improved by 1.303 and 0.032, compared to the sub-optimal result on the FFHQ. The above results demonstrate the superiority of our method in coordinated face inpainting, especially with large masked regions.

Table 1: Quantitative comparisons with SOTA methods on CelebA-HQ and FFHQ datasets. (↓ Lower is better. ↑ Higher is better)

Methods	Dataset	*L1*(%) ↓			FID ↓			PSNR ↑			SSIM ↑		
		1-20%	20-40%	40-60%	1-20%	20-40%	40-60%	1-20%	20-40%	40-60%	1-20%	20-40%	40-60%
PConv []	CelebA-HQ []	1.131	2.311	4.363	12.716	27.957	42.594	32.240	26.085	21.900	0.941	0.862	0.762
DeepFillv2 []		0.788	2.066	3.968	9.766	22.793	29.243	32.700	25.998	21.943	0.944	0.848	0.736
PIC []		0.780	2.036	4.311	4.190	11.035	21.360	33.006	25.961	21.263	0.951	0.859	0.730
CTSDG []		1.350	2.213	3.900	9.171	14.324	22.889	32.198	26.823	22.490	0.927	0.856	0.747
DSI []		0.820	2.077	4.149	9.037	20.327	29.040	32.699	26.107	21.708	0.938	0.831	0.704
ICT []		0.949	2.004	3.901	3.136	8.715	16.747	33.416	26.639	22.013	0.959	0.879	0.765
Ours		**0.585**	**1.451**	**2.937**	**2.369**	**6.410**	**12.086**	**35.772**	**28.669**	**24.177**	**0.968**	**0.900**	**0.800**
PConv []	FFHQ []	0.720	2.178	4.411	12.208	30.403	45.709	32.592	25.422	21.237	0.955	0.867	0.761
DeepFillv2 []		0.715	2.104	4.250	12.062	29.276	40.295	32.428	25.470	21.301	0.946	0.845	0.725
PIC []		0.709	2.099	4.573	5.411	14.344	27.334	32.640	25.490	20.819	0.952	0.854	0.719
CTSDG []		**0.419**	1.532	3.569	3.916	13.477	28.495	34.946	27.044	22.272	0.968	0.888	0.765
DSI []		0.746	2.067	4.340	10.483	25.772	39.127	32.659	25.780	21.241	0.941	0.834	0.702
ICT []		0.982	2.085	4.036	3.244	8.360	14.149	33.172	26.373	21.809	0.959	0.877	0.762
Ours		0.470	**1.395**	**3.068**	**2.473**	**7.170**	**13.748**	**36.046**	**28.333**	**23.575**	**0.972**	**0.903**	**0.797**

4.3 Ablation Study

In this section, we perform extensive experiments to verify the effectiveness of each module and loss in our model. Then we conduct memory design ablation analysis. All the ablation experiments are performed on the CelebA-HQ dataset. Meanwhile, the ablation study of memory design is given in supplementary materials.

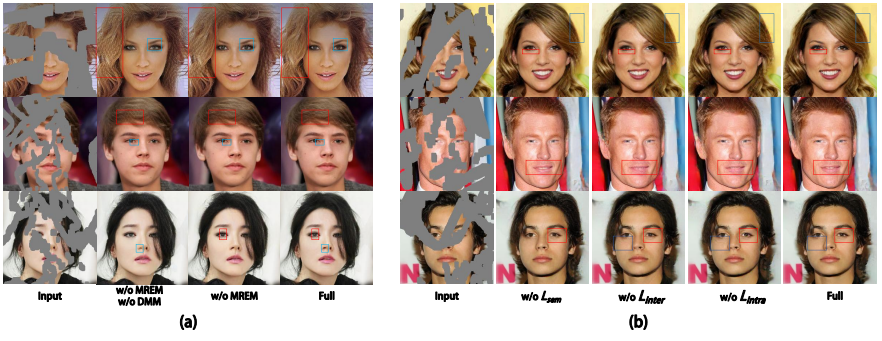

Figure 6: The qualitative comparisons result. (a) The qualitative comparisons of module ablation; (b) The qualitative comparisons of loss ablation. Zoom in for better details.

Module Ablation. We further perform module ablation to demonstrate the effectiveness of each module. There are three models with different settings for experimental comparison: 1). **w/o MREM + w/o DMM.** This model removes the MREM and DMM. The features extracted by the style encoder are injected into the Guided Refinement Network directly. 2). **w/o MREM.** This model removes the MREM and uses the fusion features after DMM to inject. 3). **Full.** Our proposed modules are all used in experiments. The module ablation results are shown in Table 2. Figure 6 (a) also shows qualitative comparisons of module ablation. The **w/o MREM + w/o DMM.** model is difficult to recover detailed textures at the semantic level, especially when some semantic categories are completely masked (e.g. the eyes of the first person in Figure 6 (a) are more blurred than the **Full** model.) Meanwhile, the **w/o MREM** recovers some detailed textures, but the faces suffer from coordination issues. In addition, the **Full** model achieved satisfactory results both in detailed textures and coordination. Finally, the **Full** model achieves the best performance. The above experimental results demonstrate that all our proposed modules are effective.

Loss Ablation. We conduct the loss ablation experiments to demonstrate the effectiveness of $\mathcal{L}_{sem}$ loss and the proposed $\mathcal{L}_{InCo^2}$ loss. The quantitative results of loss ablation are

Table 2: The evaluation results of ablation study.

Metrics	Mask Ratio	Models					
		w/o MREM w/o DMM	w/o MREM	w/o $\mathcal{L}_{sem}$	w/o $\mathcal{L}_{intra}$	w/o $\mathcal{L}_{inter}$	Full
$L1(\%)\downarrow$	1-20%	0.625	0.595	0.589	0.587	0.588	**0.585**
FID $\downarrow$		2.698	**2.221**	2.400	2.472	2.421	2.369
PSNR $\uparrow$		35.118	35.467	35.676	35.646	35.639	**35.772**
SSIM $\uparrow$		0.963	0.967	0.967	0.967	0.967	**0.968**
$L1(\%)\downarrow$	20-40%	1.550	1.473	1.461	1.452	1.462	**1.451**
FID $\downarrow$		7.205	6.529	6.541	7.011	6.874	**6.410**
PSNR $\uparrow$		28.283	28.470	28.594	28.631	28.528	**28.669**
SSIM $\uparrow$		0.890	0.897	0.898	0.897	0.897	**0.900**
$L1(\%)\downarrow$	40-60%	3.063	2.959	2.954	2.939	2.961	**2.937**
FID $\downarrow$		13.737	13.359	12.367	13.721	13.292	**12.086**
PSNR $\uparrow$		24.017	24.080	24.111	24.171	24.171	**24.177**
SSIM $\uparrow$		0.789	0.798	0.796	0.795	0.794	**0.800**

shown in Table 2. The **Full** model achieves the best performance on all metrics. Meanwhile, the removal of any loss function will degrade the performance of the model integrally. Figure 6 (b) shows the qualitative results of loss ablation. The $\mathcal{L}_{sem}$ is semantic loss, which is detailed in the supplementary material. The w/o $\mathcal{L}_{sem}$ can lead to unclear semantic boundaries. Meanwhile, w/o $\mathcal{L}_{inter}$ can not maintain coordination between corrupted regions and non-corrupted regions. Furthermore, w/o $\mathcal{L}_{intra}$ causes inconsistency within the corrupted regions. The Full model could generate reasonable results.

5 CONCLUSIONS

In this paper, we propose MDRNets for coordinated face inpainting. Specifically, we propose two collaborative modules, the DMM to establish a group of disentangled memory and the MREM to enhance feature correlation. Meanwhile, we design InCo^2 Loss, a pair of similarity based losses to better improve the inter-coordination between the corrupted and non-corrupted regions and enhance the intra-coordination in corrupted regions. Extensive experiments conducted on CelebA-HQ and FFHQ datasets demonstrate the superiority of our MDRNets.

Acknowledgements. This work is supported by China Postdoctoral Science Foundation under Grant 2022M713362 and National Natural Science Foundation of China projects under Grant No.62006227.

References

[1] Connelly Barnes, Eli Shechtman, and et al. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[2] Kyunghyun Cho, Bart van Merrienboer, and et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[3] Ian Goodfellow, Jean Pouget-Abadie, and et al. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[4] Xiefan Guo, Hongyu Yang, and et al. Image inpainting via conditional texture and structure dual generation. In *IEEE International Conference on Computer Vision*, pages 14134–14143, 2021.

[5] Martin Heusel, Hubert Ramsauer, and et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Huaibo Huang, Aijing Yu, and et al. Memory oriented transfer learning for semi-supervised image deraining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7732–7741, 2021.

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[9] Phillip Isola, Jun-Yan Zhu, and et al. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[10] Justin Johnson, Alexandre Alahi, and et al. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[11] Tero Karras, Timo Aila, and et al. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[12] Tero Karras, Samuli Laine, and et al. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[13] Xiaoming Li, Guosheng Hu, and et al. Learning symmetry consistent deep cnns for face completion. *IEEE Transactions on Image Processing*, 29:7641–7655, 2020.

[14] Yijun Li, Sifei Liu, and et al. Generative face completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017.

[15] Guilin Liu, Fitsum A Reda, and et al. Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision*, pages 85–100, 2018.

[16] Hongyu Liu, Bin Jiang, and et al. Coherent semantic attention for image inpainting. In *IEEE International Conference on Computer Vision*, pages 4170–4179, 2019.

[17] Hongyu Liu, Ziyu Wan, and et al. Deflocnet: Deep image editing via flexible low-level controls. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10765–10774, 2021.

[18] Hongyu Liu, Ziyu Wan, and et al. Pd-gan: Probabilistic diverse gan for image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9371–9381, 2021.

[19] Tomas Mikolov, Martin Karafiát, and et al. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[20] Kamyar Nazeri, Eric Ng, and et al. Edgeconnect: Structure guided image inpainting using edge prediction. In *IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[21] Taesung Park, Ming-Yu Liu, and et al. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[22] Jialun Peng, Dong Liu, and et al. Generating diverse structure for image inpainting with hierarchical vq-vae. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021.

[23] Mengshi Qi, Jie Qin, and et al. Latent memory-augmented graph transformer for visual storytelling. In *ACM International Conference on Multimedia*, pages 4892–4901, 2021.

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

[25] Nick C Tang, Yueting Zhuang, and et al. Face inpainting by feature guidance. In *IEEE International Symposium on Circuits and Systems*, pages 2613–2616. IEEE, 2009.

[26] Ziyu Wan, Jingbo Zhang, and et al. High-fidelity pluralistic image completion with transformers. In *IEEE International Conference on Computer Vision*, pages 4692–4701, 2021.

[27] Junke Wang, Shaoxiang Chen, and et al. Ft-tdr: Frequency-guided transformer and top-down refinement network for blind face inpainting. *IEEE Transactions on Multimedia*, 2022.

[28] Zhou Wang, Alan C Bovik, and et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[29] Jason Weston, Sumit Chopra, and et al. Memory networks. *CoRR*, abs/1410.3916, 2015.

[30] Chaohao Xie, Shaohui Liu, and et al. Image inpainting with learnable bidirectional attention maps. In *IEEE International Conference on Computer Vision*, pages 8858–8867, 2019.

[31] Rui Xu, Minghao Guo, and et al. Texture memory-augmented deep patch-based image inpainting. *IEEE Transactions on Image Processing*, 30:9112–9124, 2021.

[32] Chao Yang, Xin Lu, and et al. High-resolution image inpainting using multi-scale neural patch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6721–6729, 2017.

[33] Shuai Yang, Zhangyang Wang, and et al. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *European Conference on Computer Vision*, pages 601–617. Springer, 2020.

[34] Seungjoo Yoo, Hyojin Bahng, and et al. Coloring with limited data: Few-shot colorization via memory augmented networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11283–11292, 2019.

[35] Changqian Yu, Jingbo Wang, and et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision*, pages 325–341, 2018.

[36] Jiahui Yu, Zhe Lin, and et al. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[37] Jiahui Yu, Zhe Lin, and et al. Free-form image inpainting with gated convolution. In *IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[38] Yingchen Yu, Fangneng Zhan, and et al. Diverse image inpainting with bidirectional and autoregressive transformers. In *ACM International Conference on Multimedia*, pages 69–78, 2021.

[39] Chuanxia Zheng, Tat-Jen Cham, and et al. Pluralistic image completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[40] Tong Zhou, Changxing Ding, and et al. Learning oracle attention for high-fidelity face completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2020.

[41] Minfeng Zhu, Pingbo Pan, and et al. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.

[42] Peihao Zhu, Rameen Abdal, and et al. Sean: Image synthesis with semantic region-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.

[43] Yue-ting Zhuang, Yu-shun Wang, and et al. Patch-guided facial image inpainting by shape propagation. *Journal of Zhejiang University-SCIENCE A*, 10(2):232–238, 2009.