Identity-Preserving Diffusion for Face Restoration

Xiaying Bai
Yuhao Yang, Wenming Yang
Shenzhen International Graduate School
Tsinghua University
Shenzhenn, China
xiayingbai@gmail.com
yangelwm@163.com

Rui Zhu
Faculty of Actuarial Science and Insurance
City, University of London
London, United Kingdom
zhu.rui@city.ac.uk

Jing-Hao Xue
Department of Statistical Science
University College London
London, United Kingdom
jinghao.xue@ucl.ac.uk

***Abstract*—Face restoration is a critical task in computer vision, aiming to restore high-quality facial images from degraded inputs. In existing diffusion models, identity information is not well preserved when confronted with severely degradation. To address this challenge, we propose a Local Patch-Based Identity-Preserving Diffusion (LPIP-Diff) framework. Our local patch-based strategy leverages the interrelationships between neighboring patches to model highly structured facial context, which facilitates the restoration of fine-grained details and the preservation of identity-related features. We also introduce a fusion degradation estimation method that makes each overlapping area restored multiple times by adjacent patches, effectively restoring local details. The experimental results of LPIP-Diff on three publicly available datasets, including one severely degraded dataset, consistently demonstrate its superiority over the state-of-the-art methods in terms of both quantitative and qualitative evaluations, strikes a good balance between realism and fidelity, and enhances robustness against degradation.**

***Index Terms*—Diffusion model, face restoration, identity preservation, local patches.**

I. Introduction

The process of restoring high-quality (HQ) facial images from their corresponding low-quality (LQ) versions, commonly known as face restoration [1], is widely applied in various domains, including low-resolution face recognition [2]–[4] and surveillance image restoration [5], [6]. Face restoration is a challenging and ill-posed task, because there exists multiple mappings between LQ and HQ images. Severe degradations, such as downsampling, noise, blurring and compression artifacts, usually result in a deterioration of fine-grained facial details. Without such details, the restoration process is unable to accurately capture and restore the unique features that define a person's identity, which leads to the problem of loss of identity in the restored HQ image.

To address the challenge of identity loss, previous work [9] integrates the identity features into convolutional neural networks (CNN) by incorporating transformation parameters. Recently, diffusion models [10] have shown strong generative capabilities in various tasks, including image inpainting, compression artifact removal [11], image stylization [12], and graffiti editing [13]. Diffusion models, as parameterized Markov chains, have also been utilised in face restoration [7], [8], [14]. A visual comparison of some existing methods can be seen in Fig. 1. DifFace[7] proposed a novel method for blind facial restoration which approximates the true posertior by combing a transition distribution and a pre-trained diffusion model. However, such a method would introduce incorrect glasses attributes while losing face identity since the data used for pretraining is quiet different from the facial ones. DiffIR[8] utilizes diffusion model to generate a compact prior to guide the main transformer framework. Even though it performs well on several low-level vision tasks, it fails to make full use of diffusion model's generative ability. How to help diffusion model dig more realistic details without producing redundant artifacts has long been a challenge for this field.

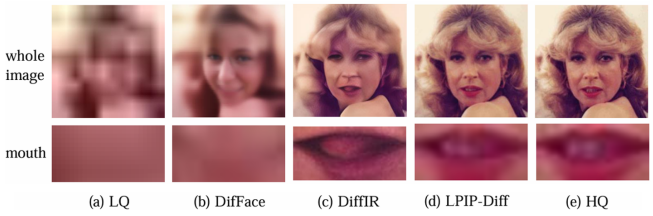


Fig. 1. Comparison of restoration results between existing diffusion restoration methods DifFace [7] and DiffIR [8] and our LPIP-Diff. The restoration results of the whole image show the identity information, and the restoration results of the local structures show how these methods restore details. Our method can capture excellent details and preserve the identity.

In this paper, we propose a local patch-based identity-preserving diffusion (LPIP-Diff) framework to accurately restore HQ facial images with fine-grained details, which assists to preserve personal identities. Specifically, we tailor the conditional diffusion model [15] via a local patch-based strategy to restore the overlapping patches of a facial image. With such strategy, each overlapping region in the HQ facial image can be restored several times via the adjacent patches. This enables us to better exploit the interrelationships between the neighbouring areas of the facial images. To this end, we propose a simple fusion degradation estimation scheme to restore the overlapping region as the average of several versions restored by the adjacent patches. In such a manner, LPIP-Diff enhances the local details of the restored HQ facial image and preserve the global identity information. Our contributions can be summarized as follows:

- We propose a local patch-based identity-preserving dif-

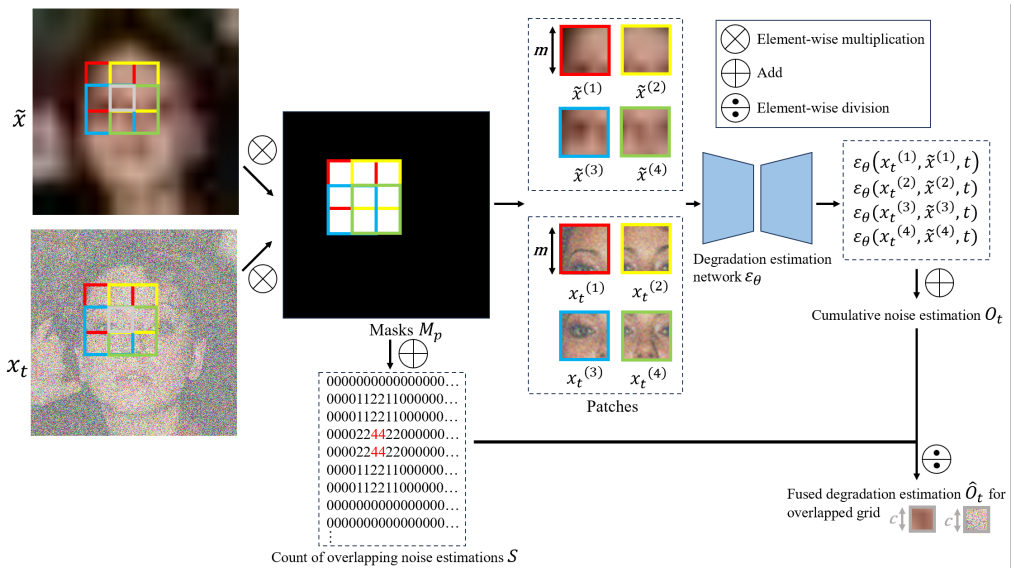


Fig. 2. An illustration of the fusion degradation estimation process.

fusion (LPIP-Diff) framework to accurately restore the fine-grained details of face images and preserve identity. To the best of our knowledge, this is the first work to introduce the idea of local patch in the diffusion model for face restoration.

- We propose a fusion degradation estimation to restore overlapping patches by leveraging the interrelationships between adjacent regions.
- We showcase the effectiveness and generalizability of LPIP-Diff through extensive experiments on three publicly available datasets, including one dataset suffering from severe degradation.

II. Methodology

A. The conditional diffusion model

The diffusion model [10], [16] aims to learn the reverse process $p_\theta(\mathbf{x}_{0:T})$, which is represented as a Markov chain with learned Gaussian transitions. The transition kernel $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ specifies the transition from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{1}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{2}$$

The forward process of the diffusion model approximates the posterior distribution $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ by incorporating it into a Markov chain and iteratively applying the diffusion kernel $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ to the observed data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. This is done by gradually introducing Gaussian noise to $\mathbf{x}_0$ based on the variance parameters $\beta_1, ..., \beta_T$, which are set as constants during training [10]:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{3}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{4}$$

Based on the above diffusion model, the conditional diffusion model [15] keeps the same forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ and learns the conditional reverse process $p_\theta(\mathbf{x}_{0:T}|\tilde{\mathbf{x}})$ where the input data distribution is $(\mathbf{x}_0, \tilde{\mathbf{x}}) \sim q(\mathbf{x}_0, \tilde{\mathbf{x}})$:

$$p_\theta(\mathbf{x}_{0:T}|\tilde{\mathbf{x}}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \tilde{\mathbf{x}}). \tag{5}$$

Since the variances of both forward and reverse processes are fixed, the conditional diffusion model aims to train the reverse process with the mean estimator $\mu_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t)$ to predict the posterior mean $\tilde{\mu}_t(\mathbf{x}_t, \tilde{\mathbf{x}}, \mathbf{x}_0)$ of the forward process:

$$\mu_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t)), \tag{6}$$

where ϵ_θ is the noise prediction network, $\alpha_t = 1-\beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. In the end, by optimizing the following reweighted simplified training objective

$$E_{t,\mathbf{x}_0,\tilde{\mathbf{x}},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \tilde{\mathbf{x}}, t)||^2], \tag{7}$$

we can obtain $\epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t)$.

Finally, in the sampling process of the conditional diffusion model, $\mathbf{x}_{t-1}$ is sampled from $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \tilde{\mathbf{x}})$:

$$\mathbf{x}_{t-1} = \sqrt{\overline{\alpha}_{t-1}}(\frac{\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t)}{\sqrt{\overline{\alpha}_t}}) + \sqrt{1-\overline{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t). \tag{8}$$

B. The LPIP-Diff framework

The framework of the proposed LPIP-Diff is depiced in Fig. 2. The main difference between LPIP-Diff and other diffusion model based methods is the utilization of Local Patch-based strategy. In Local Patch-based strategy, we choose to crop the input LQ facial image $\tilde{\mathbf{x}}$ into P overlapping patches and input only one patch into the conditional diffusion model as condition for one time instead of the whole image. Thus

we can obtain the estimated noise for every single patch in an iterative behavior. In order to obtain the estimation for the whole image, we use a Fusion Degradation Estimation scheme. Since the cropped patches are overlapping, the overlapping area may receive several different estimations from different patches. We take the average of the estimated noise as final results for such area.

Algorithm 1 The training process of LPIP-Diff

Require: High-quality images $\mathbf{x}_0$, degraded low-quality images $\tilde{\mathbf{x}}$
1: **while** Not converged **do**
2: $(\mathbf{x}_0, \tilde{\mathbf{x}}) \sim q(\mathbf{x}_0, \tilde{\mathbf{x}})$
3: $\mathbf{x}_0^{(p)} = Crop(\mathbf{x}_0)$
4: $\tilde{\mathbf{x}}^{(p)} = Crop(\tilde{\mathbf{x}})$
5: $t \sim Uniform(\{1, ..., T\})$
6: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
7: Perform gradient descent on a single patch
8: $\nabla_\theta ||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}\mathbf{x}_0^{(p)} + \sqrt{1-\overline{\alpha}_t}\epsilon, \tilde{\mathbf{x}}^{(p)}, t)||^2$
9: **end while**
Ensure: ϵ_θ

Algorithm 2 The sampling process of LPIP-Diff

Require: Degraded low-quality images $\tilde{\mathbf{x}}$, conditional diffusion model $\epsilon_\theta(\mathbf{x}_t, \tilde{\mathbf{x}}, t)$, sampling steps T, P patches
1: $\mathbf{x}_T \leftarrow$ sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, ..., 1$ **do**
3: $O_t = 0$
4: **for** $p = 1, ..., P$ **do**
5: $\mathbf{x}_t^{(p)} = Crop(\mathbf{x}_t)$
6: $\tilde{\mathbf{x}}^{(p)} = Crop(\tilde{\mathbf{x}})$
7: $O_t = O_t + \epsilon_\theta(\mathbf{x}_t^{(p)}, \tilde{\mathbf{x}}^{(p)}, t)$
8: **end for**
9: $\hat{O}_t = avg(O_t)$
10: $\mathbf{x}_{t-1} \leftarrow \sqrt{\overline{\alpha}_{t-1}}(\frac{\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t}\hat{O}_t}{\sqrt{\overline{\alpha}_t}}) + \sqrt{1-\overline{\alpha}_{t-1}}\hat{O}_t$
11: **end for**
Ensure: $\mathbf{x}_0$

A more mathematical description of our LPIP-Diff's training and sampling process is shown in Algorithms 1 and 2 respectively, in which $Crop(\cdot)$ means crop operation. The training process of LPIP-Diff doesn't differ from standard conditional diffusion model very much. The only difference is cropping operation becomes essential in LPIP-Diff. In sampling process, we firstly crop x_t and $\tilde{\mathbf{x}}$ into several overlapping patches$(\mathbf{x}_t^{(p)}, \tilde{\mathbf{x}}^{(p)})$, jointly. Then we input $(\mathbf{x}_t^{(p)}, \tilde{\mathbf{x}}^{(p)})$ and t into the degradation estimation network to obtain corresponding noise $\epsilon_\theta(\mathbf{x}_t^{(p)}, \tilde{\mathbf{x}}^{(p)}, t)$. Finally, we accumulate all the noise and take the average of them to get the fused degradation estimation O_t as the estimated noise for the whole image at timestep t. Fig. 2 provides a visual process of how we get red, yellow, blue and green four different and overlapping patches. In such occasion, the patchsize is set to 4 pixels and step size is 2 pixels. It's obvious that the middle area receives 4 different estimations from 4 adjacent patches, thus the estimation from the middle area should be the average of the 4 overlapping patches.

III. Experiments

A. Experiment settings

We trained our model on the FFHQ dataset [18] and generated LQ images according to the degradation model described in Wang et al. [19]. To demonstrate the superiority of our method, we conducted evaluations on CelebA-Test, LFW and Helen [20] datasets. Image quality assessement metrics including PSNR, SSIM [21] and ID Sim [22] The patch size m for different SR scales, $\times 4, \times 8, \times 16$ is set to 16, 32, 64 respectively, and the step size c is $\frac{1}{2}$ of the patch size.

B. Comparison with the state-of-the-arts

We performed a comprehensive quantitative and qualitative comparison between LPIP-Diff and five state-of-the-art image restoration methods, GDSSR [17], TIIN [9], SR3 [14], DifFace [7] and DiffIR [8].

1) The quantitative comparison: The quantitative metrics are shown in Table I. Obviously, whether on severely degraded or general datasets, LPIP-Diff consistently outperforms other methods in all compared metrics at different scales.

2) The qualitative comparison: The qualitative visualisations are presented in Fig. 3. Other diffusion model based methods, GDSSR, DifFace and SR3, wrongly utilized the global information and caused oversmoothed or noisy background. In comparison, our LPIP-Diff which focus more on local content can generate realistic details without introducing unpleasant artifacts.

C. Ablation studies

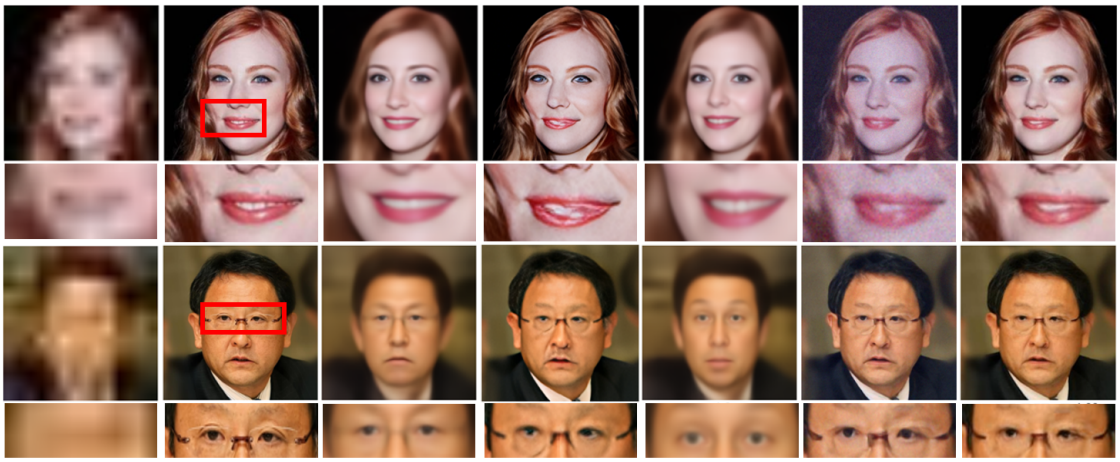
To further demonstrate the effectiveness of LPIP-Diff, we conduct ablation studies on patch size and fusion strategy.

1) The impact of patch size: To determine the best patch size, we conducted a comparative analysis on the Helen dataset, , as shown in Tables II. We employed the degradation model [19] to synthesize LQ images from HQ images of sizes 64×64, 128×128, and 256×256. The 16×16, 32×32 and 64×64 patch sizes demonstrate better results respectively. These sizes are more suitable for capturing facial structures such as eyes, mouth in the corresponding image sizes, which facilitates our diffusion model in accurately estimating the degradation of local patches and effectively merging inter-patch connections.

2) The impact of fusion strategy: We compare two fusion strategies on the Helen dataset, taking the average and median of the degradation estimates of overlapping local adjacent patches, as shown in Table III. For the magnification factors of $\times 4$ and $\times 8$, the average fusion strategy is better than the median fusion strategy, because averaging adjacent patches simultaneously considers the information of all estimates of several local patches, while the median strategy only uses the median estimate. Apparently, the former considers more of the dependencies between adjacent regions.

TABLE I
THE QUANTITATIVE EVALUATION RESULTS ON THE CELEBA-TEST, LFW AND HELEN DATASETS.

Methods	Scale	CelebA-Test			LFW			Helen		
		PSNR↑	SSIM↑	ID Sim↑	PSNR↑	SSIM↑	ID Sim↑	PSNR↑	SSIM↑	ID Sim↑
GDSSR [17]	4×	27.29	0.736	0.675	32.64	0.838	0.734	29.33	0.821	0.704
TIIN [9]		31.22	0.864	0.883	35.60	0.941	0.886	32.19	0.913	0.895
SR3 [14]		22.99	0.816	0.934	29.15	0.898	0.905	26.02	0.830	0.928
DifFace [7]		27.05	0.713	0.546	32.37	0.820	0.665	29.18	0.802	0.621
DiffIR [8]		29.89	0.834	0.815	34.28	0.913	0.827	31.21	0.852	0.824
LPIP-Diff (Ours)		**31.83**	**0.898**	**0.948**	**36.08**	**0.963**	**0.921**	**32.60**	**0.939**	**0.935**
GDSSR [17]	8×	24.39	0.665	0.691	27.69	0.807	0.593	26.02	0.789	0.618
TIIN [9]		26.59	0.718	0.796	29.18	0.840	0.675	27.55	0.824	0.697
SR3 [14]		21.85	0.672	0.827	25.41	0.812	0.705	23.95	0.798	0.730
DifFace [7]		24.23	0.659	0.643	28.02	0.798	0.552	27.34	0.779	0.569
DiffIR [8]		25.87	0.688	0.763	28.56	0.825	0.632	27.10	0.809	0.671
LPIP-Diff (Ours)		**27.16**	**0.751**	**0.846**	**29.62**	**0.868**	**0.720**	**27.97**	**0.849**	**0.738**
GDSSR [17]	16×	20.82	0.503	0.291	25.19	0.723	0.550	23.59	0.704	0.438
TIIN [9]		22.58	0.617	0.475	27.55	0.824	0.697	26.02	0.789	0.618
SR3 [14]		18.29	0.497	0.521	21.98	0.715	0.713	20.45	0.695	0.650
DifFace [7]		20.62	0.313	0.170	24.96	0.707	0.488	23.46	0.687	0.362
DiffIR [8]		21.97	0.539	0.418	26.83	0.746	0.641	25.11	0.728	0.569
LPIP-Diff (Ours)		**23.10**	**0.649**	**0.533**	**27.69**	**0.839**	**0.728**	**26.37**	**0.809**	**0.655**



(a) LQ (b) HQ (c) GDSSR (d) TIIN (e) DifFace (f) SR3 (g) LPIP-Diff

Fig. 3. The qualitative comparison between LPIP-Diff and the state-of-the-arts on the CelebA-Test dataset (8×).

TABLE II
THE COMPARISON OF DIFFERENT PATCH SIZES ON HELEN.

Patch Size	Step Size	Scale	PSNR↑	SSIM↑	ID Sim↑
4×4	2	×4	31.76	0.881	0.842
8×8	4		32.18	0.905	0.893
16×16	8		**32.60**	**0.939**	**0.935**
32×32	16		32.29	0.914	0.902
8×8	4	×8	27.10	0.793	0.631
16×16	8		27.53	0.816	0.684
32×32	16		**27.97**	**0.849**	**0.738**
64×64	32		27.68	0.822	0.699

TABLE III
THE COMPARISON OF DIFFERENT FUSION STRATEGIES ON HELEN.

fusion strategy	Scale	PSNR↑	SSIM↑	ID Sim↑
Median	×4	32.25	0.882	0.847
Average		**32.60**	**0.939**	**0.935**
Median	×8	27.42	0.805	0.659
Average		**27.97**	**0.849**	**0.738**

IV. CONCLUSION

In this work, in order to alleviate identity loss in face restoration, we present the LPIP-Diff framework, which leverages the concept of local patches to model the interrelationships between neighboring regions. We also propose a fusion degradation estimation method enabling to recover better local details. In the future, we will explore semantic-level local strategies to facilitate face restoration.

ACKNOWLEDGMENTS

This work was partly supported by the National Natural Science Foundation of China (Nos.62171251 & 62311530100) and the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen (No.KJZD20231023094700001) .

REFERENCES

[1] T. Wang, K. Zhang, X. Chen, W. Luo, J. Deng, T. Lu, X. Cao, W. Liu, H. Li, and S. Zafeiriou, "A survey of deep face restoration: Denoise, super-resolution, deblur, artifact removal," *arXiv preprint arXiv:2211.02831*, 2022.

[2] X. Yin, Y. Tai, Y. Huang, and X. Liu, "FAN: Feature adaptation network for surveillance face recognition and normalization," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[3] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 183–198.

[4] C.-Y. Low, A. B.-J. Teoh, and J. Park, "MIND-Net: A deep mutual information distillation network for realistic low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 354–358, 2021.

[5] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R. W. Lau, and M.-H. Yang, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2521–2529.

[6] G. Gao, Z. Xu, J. Li, J. Yang, T. Zeng, and G.-J. Qi, "CTCNet: A CNN-transformer cooperation network for face image super-resolution," *IEEE Transactions on Image Processing*, vol. 32, pp. 1978–1991, 2023.

[7] Z. Yue and C. C. Loy, "DifFace: Blind face restoration with diffused error contraction," *arXiv preprint arXiv:2212.06512*, 2022.

[8] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," *arXiv preprint arXiv:2303.09472*, 2023.

[9] Q. Bao, R. Zhu, B. Gang, P. Zhao, W. Yang, and Q. Liao, "Distilling resolution-robust identity knowledge for texture-enhanced face hallucination," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6727–6736.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[11] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[12] N. Huang, Y. Zhang, F. Tang, C. Ma, H. Huang, W. Dong, and C. Xu, "Diffstyler: Controllable dual diffusion for text-driven image stylization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2024.

[13] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," *arXiv preprint arXiv:2108.02938*, 2021.

[14] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[16] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[17] Y. Chi, W. Yang, and Y. Tian, "GDSSR: Toward real-world ultra-high-resolution image super-resolution," *IEEE Signal Processing Letters*, vol. 30, pp. 95–99, 2023.

[18] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[19] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9168–9178.

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*. Springer, 2012, pp. 679–692.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] X. Li, S. Zhang, S. Zhou, L. Zhang, and W. Zuo, "Learning dual memory dictionaries for blind face restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5904–5917, 2022.