MetaF2N: Blind Image Super-Resolution by Learning Efficient Model Adaptation from Faces

Zhicun Yin[1] Ming Liu[1](✉) Xiaoming Li[1] Hui Yang[2]
Longan Xiao[2] Wangmeng Zuo[1,3]
[1] Harbin Institute of Technology [2] Shanghai Transsion Co, Ltd [3] Peng Cheng Laboratory
cszcyin@outlook.com, csmliu@outlook.com, csxmli@gmail.com, wmzuo@hit.edu.cn

Abstract

Due to their highly structured characteristics, faces are easier to recover than natural scenes for blind image super-resolution. Therefore, we can extract the degradation representation of an image from the low-quality and recovered face pairs. Using the degradation representation, realistic low-quality images can then be synthesized to fine-tune the super-resolution model for the real-world low-quality image. However, such a procedure is time-consuming and laborious, and the gaps between recovered faces and the ground-truths further increase the optimization uncertainty. To facilitate efficient model adaptation towards image-specific degradations, we propose a method dubbed ***MetaF2N****, which leverages the contained* ***F****aces to fine-tune model parameters for adapting* ***to*** *the whole* ***N****atural image in a* ***Meta****-learning framework. The degradation extraction and low-quality image synthesis steps are thus circumvented in our MetaF2N, and it requires only one fine-tuning step to get decent performance. Considering the gaps between the recovered faces and ground-truths, we further deploy a MaskNet for adaptively predicting loss weights at different positions to reduce the impact of low-confidence areas. To evaluate our proposed MetaF2N, we have collected a real-world low-quality dataset with one or multiple faces in each image, and our MetaF2N achieves superior performance on both synthetic and real-world datasets. Source code, pre-trained models, and collected datasets are available at* https://github.com/yinzhicun/MetaF2N*.*

1. Introduction

With the development of dataset construction, network design, and many other relevant methods, blind image super-resolution (SR) [9, 33] has acquired enormous progress in recent years. To construct pairwise low-/high-quality samples for training the image SR models, typically one can capture low- and high-quality pairs by adjusting the focal length of cameras [5, 55, 61] and shooting distance [8], or synthesize low-quality images via degradation modeling [26, 51, 59]. However, these methods can only cover a limited and biased range of degradations, which is insufficient for real-world applications. Besides, most existing blind image SR methods [51, 59] train a static model for all testing scenarios, greatly limiting their flexibility and generalization ability.

In order to break the restriction of limited training sets, self-supervised learning has been introduced to train a model for each low-quality image, without requiring pairwise ground-truths [10, 42, 47]. Although these methods exhibit a great deal of flexibility, they largely rely on certain priors or assumptions, showing inferior image SR performance. Recently, Li *et al*. [26] reached a better compromise on the requirement of ground-truths and proposed ReDegNet by leveraging the faces contained in natural images. In specific, the face regions in a real-world low-quality natural image are processed via blind face SR methods [36, 50, 57, 63], which have achieved appealing results thanks to the highly structured characteristics of faces. Then, the low-quality and recovered face pairs can be utilized to model the degradations in the image. Finally, more low-quality images are synthesized with the degradation representations for fine-tuning the SR model. With the above design, ReDegNet [26] is flexible to process a single image or a batch of images with diverse degradations.

Although ReDegNet [26] achieves superior SR performance, especially in specific scenarios (*i.e*., fine-tuning with faces within the test image), there are still several limitations. On the one hand, the training procedure is time-consuming and computationally intensive, where the modules for degradation representation extraction and low-quality image synthesis are jointly optimized. Furthermore, when the model is intended to deal with a single low-quality image, it is difficult to determine when to terminate the training process for avoiding under-fitting or over-fitting.

On the other hand, even though the faces processed by the blind face restoration methods show appealing quality, there inevitably exist gaps between the recovered faces and the real ground-truths. Such gaps may affect the degradation representation accuracy and further hamper the training of the SR model. These problems have a large impact on the SR performance, especially when the model is fine-tuned for image-specific super-resolution.

To remedy the aforementioned problems regarding training efforts and data gaps, we propose an efficient and effective method dubbed MetaF2N. In specific, the SR model directly fine-tunes the parameters from the low-quality and recovered face pairs, then uses the fine-tuned parameters to process the whole natural image. As such, the cumbersome degradation extraction and low-quality image synthesis steps are circumvented in our method, which avoids the interference of degradation modeling errors. In order to stabilize and accelerate the fine-tuning process during inference, we adopt the model-agnostic meta-learning [11] framework in the training phase, resulting in a much more practical model adaptation procedure that can be finished in a single fine-tuning step.

Additionally, considering the gaps between low-quality faces and that processed by blind face restoration methods, we argue that treating all pixels equally will magnify the effect of the errors. Therefore, we further deploy a MaskNet, which adaptively predicts loss weights for different positions, to fulfill that the recovered face regions more similar to the ground-truths are assigned with higher loss weights. Ideally, the weight map should be extracted from the recovered faces and the ground-truths. Unfortunately, the ground-truths are unavailable during inference. Considering that predicting the loss weight map is similar to the image quality assessment (IQA) task, following the idea of degraded-reference IQA methods [2, 62], the MaskNet instead takes the low-quality and recovered faces as input. According to our observation, typically the areas closer to the ground-truth are assigned with larger weights and vice versa, which is consistent with the intuitions.

For evaluating the proposed MetaF2N, we have constructed several synthetic datasets based on FFHQ [21] and CelebA [34]. To further show the effectiveness under real-world scenarios, we have also collected a real-world low-quality image dataset from the Internet and existing datasets, namely RealFaces200 (RF200), which contains a single face or multiple faces in each image. Extensive experiments show that our MetaF2N achieves superior performance on both synthetic and real-world datasets.

In summary, the contribution of this paper includes,

- We propose an efficient and effective method dubbed MetaF2N for blind image super-resolution, which takes advances of blind face restoration models and learns model adaptation from the face regions with only one fine-tuning step.
- Considering that the gaps between faces recovered by blind face restoration methods and the ground-truth ones may result in an inaccurate degradation modeling, a MaskNet is introduced for confidence prediction to mitigate the effect of the gaps.
- A real-world low-quality image dataset with one or multiple faces in each image is collected, which will be helpful for face-guided blind image super-resolution.

2. Related Work

Recent efforts on blind image SR are mainly devoted to network design and data construction. We recommend [9, 33] for a comprehensive review of blind image SR, and focus on the most relevant methods of data construction, which are orthogonal to network design. We also briefly review recent advances in blind face restoration and meta-learning, which settle the foundation of this work.

2.1. Blind Image Super-Resolution

Pairwise Data. Recent methods have made great efforts towards pairwise datasets that cover a wider range of real-world degradations and are better aligned. The most intuitive way is to collect pairwise samples. For example, Chen *et al.* [8] collected a City100 dataset by a DSLR and a smartphone via focal length and shooting distance adjustment, respectively. Since City100 was captured via postcards ignoring the geometries, Cai *et al.* [5] and Wei *et al.* [55] proposed to capture image pairs in real-world scenes and collected two datasets RealSR and DRealSR. Joze *et al.* [20] instead captured a dataset ImagePairs by embedding two cameras with different resolutions into a beam splitter. Another approach is to synthesize low-quality images from high-quality ones. Early methods [15, 18, 19, 53] utilized classical degradation models composed of blur, down-sampling, noise, *etc*., which are insufficient to model the real-world degradations. Zhang *et al.* [59] and Wang *et al.* [51] proposed to mimic real-world degradations via random combinations of different degradations. Albeit their progress on general image SR, these methods still cover a limited and biased range of real-world degradations.

Unpaired Data. For better utilizing unpaired low-quality images that are easier to collect, some methods [12, 49, 58] proposed to extract the degradation representations or directly learn low-quality image synthesis in an unsupervised manner. Some methods [42, 47] utilized the texture recurrence across different scales or the priors embraced in the model learning process, and achieved image-specific blind image super-resolution in a self-supervised scheme. Purely relying on the unpaired samples or solely the low-quality ones, these methods may cover more degradations yet have unsatisfactory performance in most cases. As a compromise, Li *et al*. [26] utilized the highly structured faces in real-world low-quality images, and extracted degradation representations from the low-quality faces and their counterparts recovered by blind face restoration methods. In this work, we take a step forward to solve the problems of Li *et al*. [26] and propose a MetaF2N framework for effective blind image SR of LR natural images containing faces.

2.2. Blind Face Restoration

The diverse structures and complicated degradation jointly exacerbate the difficulties of blind natural image SR. In contrast, recent methods tend to leverage the structure information of specific images (*e.g*., faces [27–30, 57, 63] and text [31]), which have exhibited superior performance. Li *et al*. [28, 29] suggested using high-quality image(s) of the same person to provide personalized guidance. More general face restoration methods focused on constructing dictionaries [16, 27, 30, 54], semantic maps [7], or codebooks [63]. Recently, the generative structure prior based methods [6, 50, 57, 64] leveraged the representative pretrained StyleGAN models [22] and showed tremendous improvement in restoring fine-grained textures. Compared with natural images, these methods showed great generalization abilities, even in the presence of unknown degradations. This property has motivated us to employ face images on the inner loop of meta-learning to benefit the optimization toward better natural image SR.

2.3. Image Super-Resolution with Meta-Learning

Meta-learning aims to learn adaptation to new domains or tasks by leveraging existing ones, which has demonstrated excellent generalization ability in many tasks [11, 14, 37, 40, 44, 45, 48]. Early works like MZSR [43] and MLSR [41] utilized meta-learning techniques to obtain a better initialization of model parameters that can be efficiently adapted to new low-quality images. Meta-KernelGAN [25] utilized meta-learning for efficient kernel estimation, which can be combined with non-blind SR methods. As one can see, MZSR [43] and MLSR [41] heavily rely on known degradations to construct inner loop supervisions, while Meta-KernelGAN is also restricted due to that the degradation representations formulated by kernels cannot well describe many real-world degradations. In this work, we adopt face regions as inner loop supervision and get rid of explicit degradation representations, which is more practical for real-world scenarios.

3. Preliminary

For efficiently utilizing the generalization ability of face restoration methods in blind image SR and achieving flexibility on image-specific degradations, we have designed our MetaF2N using the model-agnostic meta-learning framework (MAML) [11]. In this section, we briefly introduce the meta-learning framework based on MAML, which will be helpful for understanding our method.

Typically, a deep model f is trained by a learning objective $\mathcal{L}$, which is directly dependent on the purpose of the task (*e.g*., ℓ_1 loss for fidelity in image SR tasks), and the optimal parameter θ^* is obtained via

$$\theta^* = \arg\min_{\theta} \mathcal{L}(f(\mathbf{x};\theta), \mathbf{y}), \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are input and ground-truth. Intuitively, the objective of Eqn. (1) is *to get a θ leading to the lowest $\mathcal{L}$*.

Instead, MAML [11] aims to learn a parameter θ^*, which can be efficiently fine-tuned toward the test sample with one or several steps. Taking the one-step situation as an example, the learning objective can be written as

$$\theta^* = \arg\min_{\theta} \sum_{T_i \sim p(T)} \mathcal{L}^{T_i}(f(\mathbf{x};\theta - \alpha\frac{\partial \mathcal{L}_{in}^{T_i}}{\partial \theta}), \mathbf{y}), \tag{2}$$

where $p(T)$, $\mathcal{L}_{in}^{T_i}$, and α respectively represent the distribution (or collection) of tasks, the inner loop loss function for the task T_i, and learning rate for the inner loop. Here, we can regard $\theta - \alpha\frac{\partial \mathcal{L}_{in}^{T_i}}{\partial \theta}$ as a whole. Then, similar to Eqn. (1), the objective of Eqn. (2) can be interpreted as *to get a θ leading to the lowest $\mathcal{L}^{T_i}$ after one back-propagation step w.r.t. $\mathcal{L}_{in}^{T_i}$*. In other words, a model trained under the MAML settings can adapt to new domains or tasks with a limited number of fine-tuning steps (*i.e*., $\theta - \alpha\frac{\partial \mathcal{L}_{in}^{T_i}}{\partial \theta}$).

4. Method

In this work, we aim to obtain a blind image SR model f, which can be efficiently adapted to process the degradation of a real-world low-quality image $\mathbf{I}_{LR}$ under the guidance of its face regions (denoted by $\mathbf{I}_{LR}^{\smiley}$) and the faces processed by blind face restoration methods (denoted by $\mathbf{I}_{BFR}^{\smiley}$).

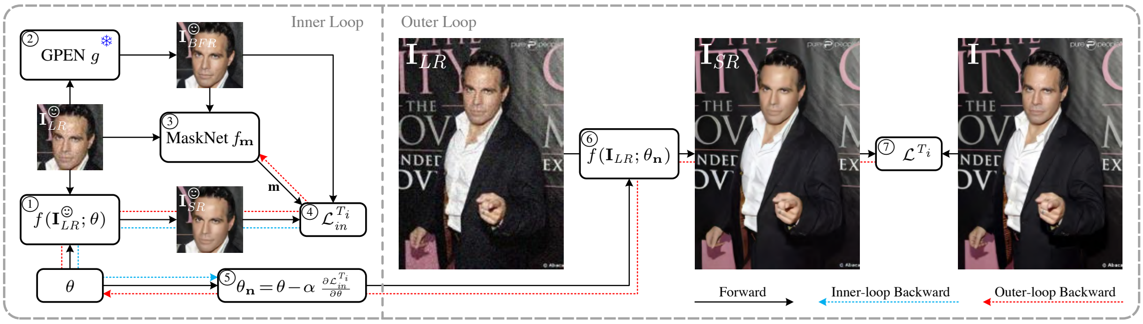
Considering the complex and wide-range degradations in real-world low-quality images, we take advantage of MAML to design our method for processing image-specific degradations. Then Eqn. (2) can be rewritten as,

$$\theta^* = \arg\min_{\theta} \sum_{T_i \sim p(T)} \mathcal{L}^{T_i}(f(\mathbf{I}_{LR};\theta - \alpha\frac{\partial \mathcal{L}_{in}^{T_i}}{\partial \theta}), \mathbf{I}), \tag{3}$$

where $p(T)$ is the distribution of different degradations, T_i is the task for one specific degradation, θ is the parameter of f and $\mathbf{I}$ denotes the ground-truth image. The design of $\mathcal{L}_{in}^{T_i}$ will be given in Sec. 4.1.

4.1. Inner Loop Design

One of the key factors of applying MAML [11] to blind image SR tasks is the design of $\mathcal{L}_{in}^{T_i}$ in Eqn. (3). Specifically, for adapting to the new degradation of an $\mathbf{I}_{LR}$, the



Inner Loop
Outer Loop
② GPEN g
$\mathbf{I}^{\smiley}_{BFR}$
$\mathbf{I}^{\smiley}_{LR}$
③ MaskNet $f_{\mathbf{m}}$
$\mathbf{m}$
① $f(\mathbf{I}^{\smiley}_{LR};\theta)$
$\mathbf{I}^{\smiley}_{SR}$
④ $\mathcal{L}^{T_i}_{in}$
θ
⑤ $\theta_{\mathbf{n}} = \theta - \alpha\frac{\partial\mathcal{L}^{T_i}_{in}}{\partial\theta}$
$\mathbf{I}_{LR}$
⑥ $f(\mathbf{I}_{LR};\theta_{\mathbf{n}})$
$\mathbf{I}_{SR}$
⑦ $\mathcal{L}^{T_i}$
$\mathbf{I}$
Forward
Inner-loop Backward
Outer-loop Backward

Figure 1: Pipeline of the proposed MetaF2N framework. During inference, the inner loop updates the initial parameter θ via a back-propagation step (dotted blue line), and the obtained θ_n is used to process the whole natural image in the outer loop. For training, the parameter update indeed relies on the gradients of outer loop loss $\mathcal{L}^{T_i}$ w.r.t. the initial parameter θ and the MaskNet parameter $\theta_{\mathbf{m}}$ (dotted red line). For easier understanding, the steps of the pipeline (remarked by circled numbers) are introduced in Sec. 4.3. Please refer to Algorithms 1 and 2 for more details about the training and inference process.

data for calculating $\mathcal{L}^{T_i}_{in}$ should have the same distribution as $\mathbf{I}_{LR}$. For MAML-based non-blind image SR methods [41, 43], one can construct inner loop data by further down-sampling $\mathbf{I}_{LR}$ with the known degradation d to get $\mathbf{I}_{LLR} = d(\mathbf{I}_{LR})$. Then assuming the inner loop loss function to be ℓ_1, we can obtain $\mathcal{L}^{T_i}_{in} = \|f(\mathbf{I}_{LLR};\theta) - \mathbf{I}_{LR}\|_1$. However, for blind image SR, it is almost impossible to get extra image pairs with the same degradation as $\mathbf{I}_{LR}$ since d is unknown.

Inspired by ReDegNet [26], which shows that faces can be utilized to extract the degradation representation for the whole image, we directly construct the inner loop data with the face regions. Thanks to the great generalization ability of blind face restoration methods [6, 50, 57, 64], we are able to obtain a pseudo ground-truth for the face regions, which is denoted by $\mathbf{I}^{\smiley}_{BFR} = f_{BFR}(\mathbf{I}^{\smiley}_{LR};\theta_{BFR})$, where f_{BFR} is the blind face restoration model with pre-trained parameter θ_{BFR}. In this paper, we follow ReDegNet [26] and adopt GPEN [57] as f_{BFR}. With the low-quality face regions $\mathbf{I}^{\smiley}_{LR}$ and the recovered $\mathbf{I}^{\smiley}_{BFR}$, the inner loop loss function can be defined by

$$\mathcal{L}^{T_i}_{in} = \|f(\mathbf{I}^{\smiley}_{LR};\theta) - \mathbf{I}^{\smiley}_{BFR}\|_1. \tag{4}$$

4.2. Adaptive Loss Weighting with MaskNet

Despite the appealing visual quality of recent blind face restoration methods, there inevitably exist gaps between generated faces $\mathbf{I}^{\smiley}_{BFR}$ and ground-truth ones $\mathbf{I}^{\smiley}$. For inaccurate regions, the degradation either explicitly (ReDegNet [26]) or implicitly (Our MetaF2N) described by the $(\mathbf{I}^{\smiley}_{LR}, \mathbf{I}^{\smiley}_{BFR})$ pair also deviate from that by $(\mathbf{I}^{\smiley}_{LR}, \mathbf{I}^{\smiley})$.

However, different regions of $\mathbf{I}^{\smiley}_{BFR}$ are treated equally in Eqn. (4), which will have negative effect on the final results. As a remedy, we propose to predict loss weights adaptively for different regions via a MaskNet. With the weight map $\mathbf{m}$ generated by the MaskNet, Eqn. (4) can be rewritten as

$$\mathcal{L}^{T_i}_{in} = \|\mathbf{m}\cdot(f(\mathbf{I}^{\smiley}_{LR};\theta) - \mathbf{I}^{\smiley}_{BFR})\|_1. \tag{5}$$

Ideally, $\mathbf{m}$ should be generated from the $(\mathbf{I}^{\smiley}_{BFR}, \mathbf{I}^{\smiley})$ pair. Unfortunately, $\mathbf{I}$ is unavailable for test samples, a possible solution is solely predicting from $\mathbf{I}^{\smiley}_{BFR}$, yet the improvement is marginal. Analogous to degraded-reference IQA [2, 62] which predicts IQA metrics from (input, output) pairs rather than (output, GT) pairs in full-reference IQA, we predict $\mathbf{m}$ from the $(\mathbf{I}^{\smiley}_{LR}, \mathbf{I}^{\smiley}_{BFR})$ pair, *i.e.*,

$$\mathbf{m} = f_{\mathbf{m}}(\mathbf{I}^{\smiley}_{LR}, \mathbf{I}^{\smiley}_{BFR};\theta_{\mathbf{m}}), \tag{6}$$

where $f_{\mathbf{m}}$ is the MaskNet, which is a simple network composed of 8 convolution layers shown in Fig. 2. As further shown in the ablation studies, the degraded-reference solution in Eqn. (6) performs better than solely predicting from $\mathbf{I}^{\smiley}_{BFR}$.

4.3. Overall Pipeline and Learning Objective

With the inner loop designed in Eqn. (5), we can build the pipeline as shown in Fig. 1. In specific, given a low-quality image $\mathbf{I}_{LR}$ containing face regions denoted by $\mathbf{I}^{\smiley}_{LR}$, we pass $\mathbf{I}^{\smiley}_{LR}$ through the SR network f and obtain the inner loop result $\mathbf{I}^{\smiley}_{SR}$①. For a fair comparison against existing methods, we take the model of Real-ESRGAN [53] as the SR network, whose parameter is used to initialize the θ in the inner loop. To construct supervision for the inner loop, we use a pre-trained GPEN [57] model, whose parameter is kept fixed in our whole pipeline②. Then we can obtain the weight map $\mathbf{m}$ following Eqn. (6)③, as well as the inner loop loss $\mathcal{L}^{T_i}_{in}$ following Eqn. (5)④, which can be used to generate a temporary parameter $\theta_{\mathbf{n}} = \theta - \alpha\frac{\partial\mathcal{L}^{T_i}_{in}}{\partial\theta}$⑤ for processing the whole natural image in the outer loop, *i.e.*,

$$\mathbf{I}_{SR} = f(\mathbf{I}_{LR};\theta_{\mathbf{n}}) = f(\mathbf{I}_{LR};\theta - \alpha\frac{\partial\mathcal{L}^{T_i}_{in}}{\partial\theta})⑥. \tag{7}$$

Algorithm 1: Training Process of MetaF2N

```
Input: Distribution of different degradation
       restoration tasks $p(T)$;
       High-quality natural image dataset $\mathcal{D}_{HR}$;
       High-quality face dataset $\mathcal{D}_{☺}$;
       Learning rates $\alpha, \beta, \gamma, \eta$;
Output: Model parameter $\theta, \theta_{\mathbf{m}}$
Initialize $f$, $D$ with $\theta$, $\theta_D$ from Real-ESRGAN [51]
Initialize $f_{BFR}$ with $\theta_{BFR}$ from GPEN [57]
Random initialize $f_{\mathbf{m}}$ with $\theta_{\mathbf{m}}$
for all training steps do
    Sample batch of tasks $T_i \sim p(T)$
    for all $T_i$ do
        Sample images $\mathbf{I}, \mathbf{I}^{☺}$ from $\mathcal{D}_{\mathcal{HR}}$ and $\mathcal{D}_{☺}$
        Create $\mathbf{I}, \mathbf{I}_{LR}, \mathbf{I}^{☺}_{LR}$ of $T_i$
        $\mathbf{I}^{☺}_{BFR} = f_{BFR}(\mathbf{I}^{☺}_{LR}; \theta_{BFR})$
        $\mathbf{m} = f_{\mathbf{m}}(\mathbf{I}^{☺}_{LR}, \mathbf{I}^{☺}_{BFR}; \theta_{\mathbf{m}})$
        for 1 step of inner loop do
            $\mathcal{L}^{T_i}_{in} = \|\mathbf{m} \cdot (f(\mathbf{I}^{☺}_{LR}; \theta) - \mathbf{I}^{☺}_{BFR})\|_1$
            $\theta_{\mathbf{n}} \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}^{T_i}_{in}(\theta)$
        end
        $\mathbf{I}_{SR} = f(\mathbf{I}_{LR}; \theta_{\mathbf{n}})$
        $\mathcal{L}^{T_i}$ is defined as Eqn. (13)
        $\mathcal{L}^{T_i}_D$ is defined as Eqn. (11)
    end
    $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} \mathcal{L}^{T_i}$
    $\theta_{\mathbf{m}} \leftarrow \theta_{\mathbf{m}} - \gamma \nabla_{\theta_{\mathbf{m}}} \sum_{T_i \sim p(T)} \mathcal{L}^{T_i}$
    $\theta_D \leftarrow \theta_D - \eta \nabla_{\theta_D} \sum_{T_i \sim p(T)} \mathcal{L}^{T_i}_D$
end
```

Algorithm 2: Inference Process of MetaF2N

```
Input: LR test image $\mathbf{I}_{LR}$;
       Trained model parameter $\theta, \theta_{\mathbf{m}}$;
       Pre-trained GPEN [57] parameter $\theta_{BFR}$;
       Fine-tuning steps $n$ and learning rate $\alpha$;
Output: Super-resolved image $\mathbf{I}_{SR}$
Initialize $f$, $f_{\mathbf{m}}$ with $\theta$, $\theta_{\mathbf{m}}$
Initialize $f_{BFR}$ with $\theta_{BFR}$ from GPEN [57]
Extract face regions $\mathbf{I}^{☺}_{LR}$ from $\mathbf{I}_{LR}$
$\mathbf{I}^{☺}_{BFR} = f_{BFR}(\mathbf{I}^{☺}_{LR}; \theta_{BFR})$
$\mathbf{m} = f_{\mathbf{m}}(\mathbf{I}^{☺}_{LR}, \mathbf{I}^{☺}_{BFR}; \theta_{\mathbf{m}})$
for n steps of inner loop do
    $\mathcal{L}_{in} = \|\mathbf{m} \cdot (f(\mathbf{I}^{☺}_{LR}; \theta) - \mathbf{I}^{☺}_{BFR})\|_1$
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{in}(\theta)$
end
$\theta_{\mathbf{n}} \leftarrow \theta$
$\mathbf{I}_{SR} = f(\mathbf{I}_{LR}; \theta_{\mathbf{n}})$
return $\mathbf{I}_{SR}$
```

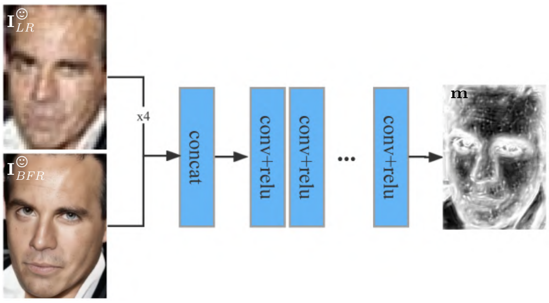


Figure 2: The network structure of MaskNet.

The above process (as remarked by the circled numbers) describes the pipeline of MetaF2N during inference and the first six steps in the training phase.

For training MetaF2N, we also need to calculate the outer loop loss⑦, which is composed of fidelity loss (ℓ_1), LPIPS loss [60], and GAN loss [13]. In specific, the fidelity loss is

$$\mathcal{L}_1 = \|\mathbf{I}_{SR} - \mathbf{I}\|_1, \tag{8}$$

and the LPIPS loss is defined by

$$\mathcal{L}_{\text{LPIPS}} = \|\phi(\mathbf{I}_{SR}) - \phi(\mathbf{I})\|_2, \tag{9}$$

where ϕ is the pre-trained AlexNet [24] feature extractor for calculating LPIPS. The adversarial loss follows the setting of Real-ESRGAN [51], which is defined by,

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}[\log(D(\mathbf{I}_{SR}))], \tag{10}$$

where the discriminator D is iteratively trained along with the SR network, *i.e.*,

$$\mathcal{L}_D = -\mathbb{E}[\log(D(\mathbf{I})) - \log(1 - D(\mathbf{I}_{SR}))]. \tag{11}$$

Note that some works [25,32] have explored training GANs under the meta-learning structure, and all of them update the discriminator loss in both the inner loop and outer loop. However, to accelerate the training process and save memory, we only update the discriminator loss in the outer loop, which also shows satisfactory results.

To improve the numerical stability and avoid gradient exploding/vanishing, we constrain the MaskNet $f_{\mathbf{m}}$ via a regularization term, which is defined by

$$\mathcal{L}_{\text{reg}} = \|\mathbf{m} - \mathbf{1}\|_2. \tag{12}$$

In summary, the learning objective of our MetaF2N (*i.e.*, the outer loop loss) is

$$\mathcal{L}^{T_i} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{LPIPS}} + \lambda_3 \mathcal{L}_{\text{adv}} + \lambda_4 \mathcal{L}_{\text{reg}}, \tag{13}$$

where the hyperparameters λ_1, λ_2, λ_3, and λ_4 are empirically set to 1, 0.5, 0.1, and 0.002, respectively. Please refer to Algorithms 1 and 2 for more details on the training and inference process of the proposed MetaF2N.

5. Experiments

5.1. Dataset and Training Details

Training Data. For synthesizing data to train the proposed MetaF2N, high-quality facial and natural images are required for inner and outer loop, respectively. It is worth noting that, the requirement on the training data is that $\mathbf{I}^{\text{☺}}_{LR}$ and $\mathbf{I}_{LR}$ share the identical degradation, but they are not necessarily from the same image. Therefore, to better leverage existing high-quality facial and natural image datasets, we use the first 30,000 images of the aligned FFHQ [21] dataset for the inner loop, and adopt DF2K[1] with 3,450 images in total for the outer loop following recent blind image SR methods [26, 51, 59]. For a fair comparison against other methods, we follow the degradation setting of Real-ESRGAN [51] for low-quality image synthesis. More details are given in the supplementary material.

Testing Data. To comprehensively evaluate the proposed MetaF2N, we have constructed several testing datasets.

Synthetic datasets are built upon the in-the-wild version of FFHQ [21] and CelebA [34], where 1,000 high-quality images are extracted from each dataset, and faces occupy around 10% areas of the whole image on average. With the high-quality images, we first construct two test datasets whose degradations are independent and identically distributed as the training set, which are denoted by FFHQ$_{iid}$ and CelebA$_{iid}$, respectively. For evaluating the generalization ability on out-of-distribution degradations, we further construct FFHQ$_{ood}$ and CelebA$_{ood}$ by changing the parameters of the degradation model, *e.g.*, Gaussian blur → motion blur, Gaussian/Poisson noise → Speckle noise. The detailed parameters are in the supplementary material.

A *real-world dataset* RealFaces200 (RF200) is also established for evaluation under a real scenario, which contains 200 real-world low-quality images collected from the Internet or existing datasets (*e.g.*, WIDER FACE [56]), and there are one or multiple faces in each image.

Implementation Details. The proposed MetaF2N mainly focuses on leveraging meta-learning for efficient model adaptation for blind image SR, which is independent of network architecture and can be incorporated into arbitrary second-order differentiable models. Therefore, we follow Real-ESRGAN [51], BSRGAN [59], and ReDegNet [26] to adopt the architecture of ESRGAN [53]. For training, each image is cropped into patches, and the patch size is 128×128 for the inner loop (*i.e.*, $\mathbf{I}^{\text{☺}}_{LR}$, $\mathbf{I}^{\text{☺}}_{BFR}$, and $\mathbf{I}^{\text{☺}}_{SR}$) and 256×256 for the outer loop (*i.e.*, $\mathbf{I}_{LR}$, $\mathbf{I}_{SR}$, and $\mathbf{I}$). The Adam [23] optimizer with β_1 = 0.5 and β_2 = 0.999 is adopted. For the SR model f, the learning rate (lr) of the inner and outer loops are 1×10^{-2} and 3×10^{-5}, respectively, while the lr for the discriminator and MaskNet are 1×10^{-4}. The MetaF2N is trained for one week and all experiments are conducted on a server with one RTX A6000 GPU.

5.2. Comparison with State-of-the-art Methods

To show the effectiveness of the proposed MetaF2N, we compare with several state-of-the-art blind image SR methods, including ESRGAN [53], RealSR [5], Real-ESRGAN [51], BSRGAN [59], MM-RealSR [39], and ReDegNet [26]. For quantitative evaluation, we utilize PSNR, LPIPS [60], FID [17], and NIQE [38] for synthetic datasets. As for the real-world dataset, since no ground-truth is available, PSNR and LPIPS are omitted. It is worth noting that KID is more accurate and appropriate than FID when evaluating the quality of images with a smaller amount of samples [3]. Thus we use KID [3] and NIQE [38] for quantitative evaluation on our RF200 dataset. To accurately calculate the distance between SR results and real-world high-quality images (*i.e.*, FID and KID), we construct these types of reference images following [4], which estimates a blurriness score based on the total variance of the Laplacian of an image. Finally, 3,808 high-quality images are selected from the in-the-wild version of FFHQ [21], whose scores are higher than the threshold defined in [4].

Quantitative Comparison. The quantitative evaluation results are provided in Tabs. 1 and 2. We provide three results of our Meta-F2N, *i.e.*, Ours (1), Ours (10), and Ours (20), where the number in the parentheses denotes the amount of fine-tuning steps during inference. All three models are initialized with the same parameter θ. For the degradations independent and identically distributed as the training set (*i.e.*, FFHQ$_{iid}$ and CelebA$_{iid}$), our MetaF2N can achieve superior performance on LPIPS, FID, and NIQE with only one fine-tuning step, and can surpass most of the competing methods on PSNR. Besides, with more fine-tuning iterations, the image quality can be further enhanced. For out-of-distribution degradations in Tab. 2 (*i.e.*, FFHQ$_{ood}$, CelebA$_{ood}$, and RF200), the trends are similar to Tab. 1, and our MetaF2N again outperforms others on most metrics, especially on the real-world RF200 dataset.

As for the competing methods, ESRGAN [53] and RealSR [5] are trained with simple degradation models, which lead to unsatisfactory performance under more complex and realistic degradations. Real-ESRGAN [51], BSRGAN [59], MM-RealSR [39], and ReDegNet [26] leverage more realistic degradation models or even real-world degradations, which contribute to much better results, yet their performance is still limited by the deterministic model and fixed degradation range. Furthermore, we also fine-tune ReDegNet [26] on RF200, and the results are provided in Tab. 2 (ReDegNet†). Compared to MetaF2N that achieves decent performance with only one fine-tuning step, marginal improvements are observed on NIQE even the ReDegNet [26] is carefully fine-tuned hundreds iterations for each image.

[1] DF2K is the combination of DIV2K [1] and Flickr2K [46] datasets.

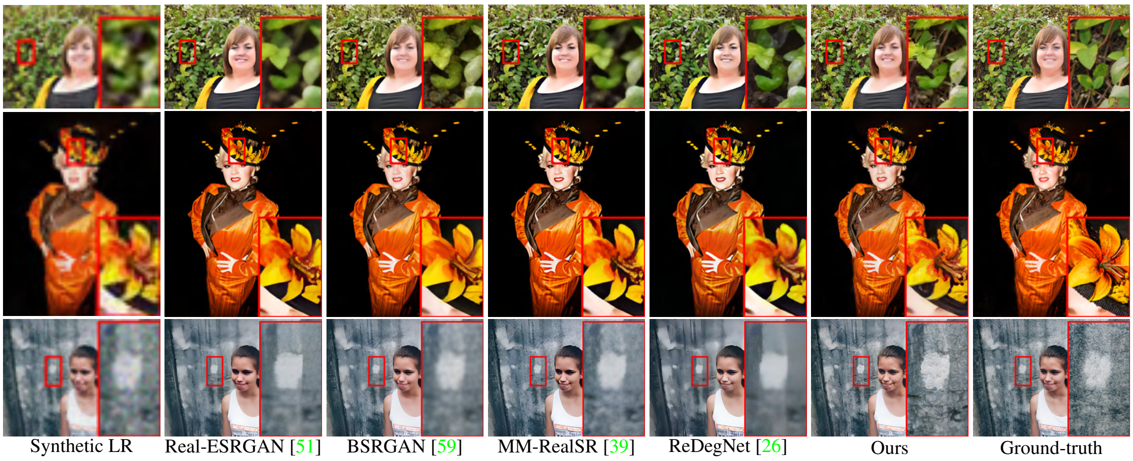


Figure 3: Visual comparison against state-of-the-art blind SR methods on synthetic datasets. Note that the results of Ours are produced by MetaF2N with one-step fine-tuning. Please zoom in for better observation.

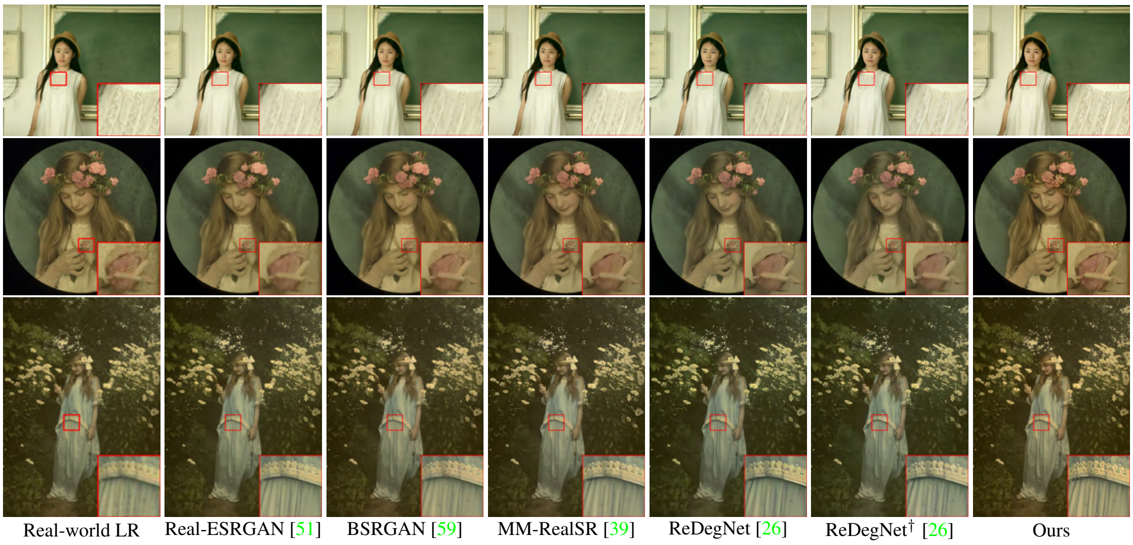


Figure 4: Visual comparison against state-of-the-art blind SR methods on real-world low-quality images in our RF200 dataset. ReDegNet† means that model is fine-tuned for each image following the official configurations of ReDegNet [26].

Qualitative Comparison. Apart from the quantitative comparison, we also compare our MetaF2N with state-of-the-art blind image SR methods qualitatively. Due to the space limit, we show the results generated by Ours (1), which is fine-tuned for 1 step with the face regions, and more qualitative results (including those generated by Ours (10) and Ours (20)) are provided in the supplementary material. As shown in Figs. 3 and 4, our results are much clearer and contain more photo-realistic textures, which can be ascribed to the effectiveness of our image-specific fine-tuning scheme.

5.3. The Visual Results of MaskNet

To better explain the effect of our MaskNet, we calculate the error map EM between ground-truth faces $\mathbf{I}^{\smiley}$ and GPEN [57] generated faces $\mathbf{I}^{\smiley}_{BFR}$ directly through subtraction and normalization, *i.e.*,

$$EM = \frac{|\mathbf{I}^{\smiley} - \mathbf{I}^{\smiley}_{BFR}|}{\max(|\mathbf{I}^{\smiley} - \mathbf{I}^{\smiley}_{BFR}|)}. \tag{14}$$

In Fig. 5, we show the predicted $\mathbf{m}$ and $1 - EM$. One can see that the predicted $\mathbf{m}$ follows similar distribution to $1 - EM$, indicating that our MaskNet has the ability to predict the gap between generated faces and real ones, which benefits the blind image SR performance. Besides, the predicted $\mathbf{m}$ is smoother than $1 - EM$, due to that 1) $\mathbf{m}$ is predicted from ($\mathbf{I}^{\smiley}_{LR}$ and $\mathbf{I}^{\smiley}_{BFR}$), which lead to less accurate results, and 2) a regularization term $\mathcal{L}_{\text{reg}} = \|\mathbf{m} - 1\|_2$ is applied to constrain $\mathbf{m}$ not to deviate too far from 1.

Table 1: Quantitative comparison on synthetic datasets with independent and identically distributed degradations as the training set. We also show the datasets used for training each method. Note that RealFaces used by ReDegNet [26] consists of 10,000 real-world low-quality face images collected by the authors. The numbers in the parentheses are the image-specific fine-tuning steps of our MetaF2N during inference. The best and second-best results are highlighted by **bold** and underline.

Methods	Training Set	FFHQ_{iid}				CelebA_{iid}			
		PSNR↑	LPIPS↓	FID↓	NIQE↓	PSNR↑	LPIPS↓	FID↓	NIQE↓
ESRGAN [53]	DF2K [1,46], OST [52]	25.10	0.642	123.67	7.56	24.78	0.555	99.10	6.27
RealSR [19]	DF2K [1,46]	25.39	0.597	122.23	7.59	24.74	0.549	99.46	5.77
Real-ESRGAN [51]	DF2K [1,46], OST [52]	**26.35**	0.293	50.70	4.85	25.81	0.311	51.69	5.00
BSRGAN [59]	DF2K [1,46], FFHQ [21], WED [35]	26.27	0.305	50.98	4.63	**25.86**	0.309	50.05	4.83
MM-RealSR [39]	DF2K [1,46], OST [52]	25.39	0.295	50.83	4.58	24.94	0.311	52.30	4.76
ReDegNet [26]	DF2K [1,46], FFHQ [21], RealFaces	25.64	0.309	57.04	4.79	25.17	0.329	56.84	4.94
Ours (1)	DF2K [1,46], FFHQ [21]	26.13	0.281	45.22	**3.81**	25.63	**0.289**	45.72	**3.97**
Ours (10)	DF2K [1,46], FFHQ [21]	26.22	0.280	44.94	3.87	25.70	**0.289**	45.43	4.03
Ours (20)	DF2K [1,46], FFHQ [21]	26.30	**0.279**	**44.51**	3.94	25.76	**0.289**	**45.22**	4.10

Table 2: Quantitative comparison on synthetic datasets with out-of-distribution degradations and the collected real-world dataset RF200. ReDegNet† means that model is fine-tuned for each image following the official configurations of ReDegNet [26]. The numbers in the parentheses are the image-specific fine-tuning steps of our MetaF2N during inference. The best and second-best results are highlighted by **bold** and underline.

Methods	FFHQ_{ood}				CelebA_{ood}				RF200	
	PSNR↑	LPIPS↓	FID↓	NIQE↓	PSNR↑	LPIPS↓	FID↓	NIQE↓	KID↓	NIQE↓
ESRGAN [53]	24.69	0.659	101.43	6.94	24.07	0.567	87.61	5.64	21.8	5.32
RealSR [19]	24.91	0.587	97.42	6.33	23.98	0.567	90.30	5.51	21.9	5.00
Real-ESRGAN [51]	**25.73**	0.302	47.87	5.34	25.07	0.322	48.77	5.43	22.4	3.82
BSRGAN [59]	**25.73**	0.298	46.40	4.78	**25.33**	0.313	48.12	4.91	22.1	4.11
MM-RealSR [39]	25.03	0.297	47.30	4.99	24.41	0.317	48.76	5.10	22.1	4.12
ReDegNet [26]	25.54	0.304	48.47	5.09	24.76	0.331	52.33	5.28	23.5	4.07
ReDegNet† [26]	-	-	-	-	-	-	-	-	23.8	3.56
Ours (1)	25.49	0.284	45.07	**4.22**	24.87	**0.297**	47.03	**4.32**	21.2	**3.03**
Ours (10)	25.57	0.284	44.64	4.30	24.93	**0.297**	46.50	4.39	21.0	3.08
Ours (20)	25.65	**0.283**	**44.30**	4.36	25.00	**0.297**	**46.23**	4.47	**20.8**	3.12

5.4. Ablation Studies

To verify the effectiveness of the components of the proposed MetaF2N, we have also conducted several ablation studies. Without loss of generality, all ablation studies are performed with FFHQ_{iid} based on Ours (1). Due to the space limit, qualitative results of ablation studies are given in the supplementary material.

MaskNet Configuration. To show the effectiveness of the MaskNet, we train a variant of our MetaF2N by removing the MaskNet. Besides, as introduced in Sec. 1, we design the MaskNet inspired by degraded-reference IQA methods, which takes both low-quality and restored faces as input (denoted by MaskNet_{DR}). For verifying the design, we also consider another variant of MetaF2N which utilizes the non-reference scheme, *i.e.*, only the restored face is taken by the MaskNet (denoted by MaskNet_{NR}). As shown in Tab. 3, the deployment of MaskNet brings considerable improvements, and utilizing the low-quality face image as a reference is also useful for our task.

Training Data Configuration. Since ground-truth faces are available during training, we train another variant by using the ground-truth face image as the inner loop supervision. Note that the MaskNet is trained together with the SR model, and we omitted it when training the version using ground-truth faces as inner loop supervision since the optimal solution should be an all-one mask. As shown in Tab. 3, the performance decreases by a large margin compared to MetaF2N, which is attributed to the gaps in the inner loop supervision between the training and inference phases.

Inner Loop Supervision Scheme. Since there are often multiple faces in a single image, as a natural extension of our method, it is intuitive to apply multiple faces to stabilize the fine-tuning process. Therefore, we also explore the inner loop supervision scheme based on another 1,000 images sampled from the in-the-wild version of FFHQ [21] with multiple faces. As shown in Tab. 4, using patches of multiple faces can slightly benefit the final performance.

Influence of Different Face Hallucination Network. To

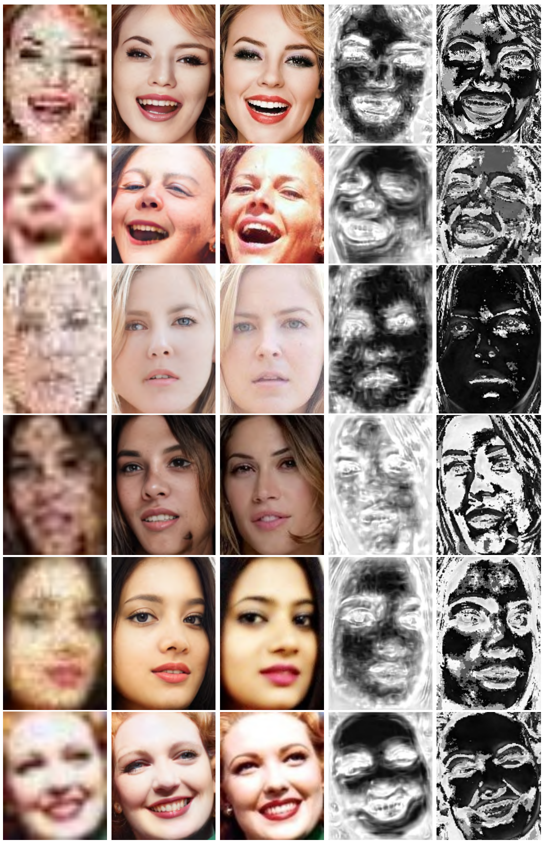


Figure 5: Visual results of MaskNet.

find the influence of different face hallucination network for the MetaF2N, we conduct an ablation study by replacing GPEN [57] with CodeFormer [63] during training and testing. As shown in Tab. 5, one can see that our MetaF2N shows satisfactory robustness with different face hallucination networks for both FFHQ$_{iid}$ and FFHQ$_{ood}$. Although the CodeFormer [63] can restore higher quaility faces compared with GPEN [57], our MaskNet minish the gap through allocating weights for each pixel of restored faces, which enhances the robustness of MetaF2N for different face hallucination networks.

Table 3: Ablation study on the training scheme regarding data configuration and the MaskNet. Train/Test Faces denote the setting of inner-loop supervision during the training and inference stages.

MaskNet	Train Faces	Test Faces	PSNR↑	LPIPS↓	FID↓	NIQE↓
-	GT	GPEN	25.68	0.285	46.96	4.08
-	GPEN	GPEN	**26.17**	0.284	47.89	4.21
MaskNet$_{NR}$	GPEN	GPEN	26.07	0.284	47.97	4.09
MaskNet$_{DR}$	GPEN	GPEN	26.13	**0.281**	**45.22**	**3.81**

Table 4: Ablation study on inner loop supervision, which is performed on 1,000 images with multiple faces (∼2.3 faces/image) from the in-the-wild version of FFHQ [21].

# of Faces	Data Form	PSNR↑	LPIPS↓	FID↓	NIQE↓
Single	Full image	25.73	0.288	42.96	**3.63**
Single	32 patches	25.77	0.288	42.89	3.65
Multiple	32 patches in total	**25.79**	**0.287**	42.85	3.65
Multiple	32 patches per face	**25.79**	**0.287**	**42.81**	3.65

Table 5: Ablation study on face hallucination networks which is performed on both FFHQ$_{iid}$ and FFHQ$_{ood}$.

Testsets	Train Faces	Test Faces	PSNR↑	LPIPS↓	FID↓	NIQE↓
FFHQ$_{iid}$	GPEN	GPEN	**26.13**	0.281	**45.22**	**3.81**
	CodeFormer	CodeFormer	26.10	**0.278**	45.83	3.83
FFHQ$_{ood}$	GPEN	GPEN	25.49	0.284	45.07	**4.22**
	CodeFormer	CodeFormer	**25.51**	**0.281**	**44.58**	4.24

5.5. Limitations and Future Work

Since our MetaF2N is trained in a MAML framework, the performance before fine-tuning is unconstrained. As such, our model requires fine-tuning with the contained faces for one or multiple steps, making it cannot be directly applied if no faces exist in real-world low-quality images, which limits the application scenarios of our method. In the future, we plan to solve this problem by collecting a real-world low-quality training set and providing a better initialization of parameters that can be used to achieve satisfactory performance even without fine-tuning.

6. Conclusion

In this paper, we proposed a MetaF2N framework for face-guided blind image SR. The MetaF2N improves the practicability of leveraging blind face restoration for processing image-specific degradations. A MaskNet is deployed to predict the loss weights for different positions considering the gaps between recovered faces and potential ground-truth ones. The proposed MetaF2N achieves superior performance on synthetic datasets with both independent and identically distributed and out-of-distribution degradations compared to training sets. The effectiveness is also validated on the collected real-world dataset.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2022YFA1004103 and the National Natural Science Foundation of China (NSFC) under Grant No. U19A2073.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 6, 8

[2] Shahrukh Athar and Zhou Wang. Degraded reference image quality assessment. *IEEE Transactions on Image Processing*, pages 822–837, 2023. 2, 4

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, pages 1–36, 2018. 6

[4] Will Brennan. Blurdetection2. https://github.com/WillBrennan/BlurDetection2, 2017. 6

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *IEEE International Conference on Computer Vision*, pages 3086–3095, 2019. 1, 2, 6

[6] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. GLEAN: Generative latent bank for large-factor image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14245–14254, 2021. 3, 4

[7] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11896–11905, 2021. 3

[8] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1652–1660, 2019. 1, 2

[9] Honggang Chen, Xiaohai He, Linbo Qing, Yuanyuan Wu, Chao Ren, Ray E Sheriff, and Ce Zhu. Real-world single image super-resolution: A brief review. *Information Fusion*, pages 124–145, 2022. 1, 2

[10] Mohammad Emad, Maurice Peemen, and Henk Corporaal. DualSR: Zero-shot dual learning for real-world super-resolution. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1630–1639, 2021. 1

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017. 2, 3

[12] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *IEEE International Conference on Computer Vision Workshops*, pages 3599–3608, 2019. 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, pages 139–144, 2020. 5

[14] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018. 3

[15] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 2

[16] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143, 2022. 3

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 6

[18] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, pages 5632–5643, 2020. 2

[19] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 466–467, 2020. 2, 8

[20] Hamid Reza Vaezi Joze, Ilya Zharkov, Karlton Powell, Carl Ringler, Luming Liang, Andy Roulston, Moshe Lutz, and Vivek Pradeep. ImagePairs: Realistic super resolution dataset via beam splitter camera rig. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 518–519, 2020. 2

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 6, 8, 9

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, pages 84–90, 2017. 5

[25] Royson Lee, Rui Li, Stylianos I Venieris, Timothy Hospedales, Ferenc Huszár, and Nicholas D Lane. Meta-learned kernel for blind super-resolution kernel estimation. *arXiv preprint arXiv:2212.07886*, 2022. 3, 5

[26] Xiaoming Li, Chaofeng Chen, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. From face to natural image: Learning real degradation for blind image super-resolution. In *European Conference on Computer Vision*, pages 376–392, 2022. 1, 3, 4, 6, 7, 8

[27] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415, 2020. 3

[28] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020. 3

[29] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *European Conference on Computer Vision*, pages 272–289, 2018. 3

[30] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2022. 3

[31] Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10103–10113, 2023. 3

[32] Weixin Liang, Zixuan Liu, and Can Liu. Dawson: A domain adaptive few shot generation framework. *arXiv preprint arXiv:2001.00576*, 2020. 5

[33] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–19, 2022. 1, 2

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 2, 6

[35] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, pages 1004–1016, 2016. 8

[36] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020. 1

[37] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017. 3

[38] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, pages 209–212, 2012. 6

[39] Chong Mou, Yanze Wu, Xintao Wang, Chao Dong, Jian Zhang, and Ying Shan. Metric learning based interactive modulation for real-world super-resolution. In *European Conference on Computer Vision*, pages 723–740, 2022. 6, 7, 8

[40] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: Task dependent adaptive metric for improved few-shot learning. *Advances in Neural Information Processing Systems*, pages 719–729, 2018. 3

[41] Seobin Park, Jinsu Yoo, Donghyeon Cho, Jiwon Kim, and Tae Hyun Kim. Fast adaptation to super-resolution networks via meta-learning. In *European Conference on Computer Vision*, pages 754–769, 2020. 3, 4

[42] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 1, 2

[43] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020. 3, 4

[44] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 3

[45] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 3

[46] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 6, 8

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 1, 2

[48] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016. 3

[49] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10581–10590, 2021. 2

[50] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 1, 3, 4

[51] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *IEEE International Conference on Computer Vision*, pages 1905–1914, 2021. 1, 2, 5, 6, 7, 8

[52] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 8

[53] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision Workshops*, pages 63–79, 2018. 2, 4, 6, 8

[54] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. RestoreFormer: High-quality blind face restoration from undegraded key-value pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17491–17500, 2022. 3

[55] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision*, pages 101–117, 2020. 1, 2

[56] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016. 6

[57] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 1, 3, 4, 5, 7, 9

[58] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 2

[59] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 2, 6, 7, 8

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5, 6

[61] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 1

[62] Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *IEEE International Conference on Computer Vision*, pages 10242–10251, 2021. 2, 4

[63] Shangchen Zhou, Kelvin CK Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*, 2022. 1, 3, 9

[64] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7662–7671, 2022. 3, 4