Super-Resolving Cross-Domain Face Miniatures by Peeking at One-Shot Exemplar

Peike Li[1,3]*, Xin Yu[1], Yi Yang[2]
[1]ReLER, Centre for Artificial Intelligence, University of Technology Sydney
[2]CCAI, College of Computer Science and Technology, Zhejiang University
[3]Baidu Research

Abstract

Conventional face super-resolution methods usually assume testing low-resolution (LR) images lie in the same domain as the training ones. Due to different lighting conditions and imaging hardware, domain gaps between training and testing images inevitably occur in many real-world scenarios. Neglecting those domain gaps would lead to inferior face super-resolution (FSR) performance. However, how to transfer a trained FSR model to a target domain efficiently and effectively has not been investigated. To tackle this problem, we develop a Domain-Aware Pyramid-based Face Super-Resolution network, named DAP-FSR network. Our DAP-FSR makes the first attempt to super-resolve LR faces from a target domain by exploiting only a pair of high-resolution (HR) and LR exemplar in the target domain. To be specific, our DAP-FSR firstly employs its encoder to extract the multi-scale latent representations of the input LR face. Considering only one target domain example is available, we propose to augment the target domain data by mixing the latent representations of the target domain face and source domain ones, and then feed the mixed representations to the decoder of our DAP-FSR. The decoder will generate new face images resembling the target domain image style. The generated HR faces in turn are used to optimize our decoder to reduce the domain gap. By iteratively updating the latent representations and our decoder, our DAP-FSR will be adapted to the target domain, thus achieving authentic and high-quality upsampled HR faces. Extensive experiments on three benchmarks validate the effectiveness and superior performance of our DAP-FSR compared to the state-of-the-art methods.

1. Introduction

Face Super-Resolution (FSR), also known as face hallucination, aims at reconstructing high-resolution (HR) face images from input low-resolution (LR) ones. FSR provides critical information for the downstream computer vision and machine learning tasks, such as face detection [1], recognition [31] and photo-editing [22, 62, 27]. Thanks to the advance of generative adversarial networks [15], FSR
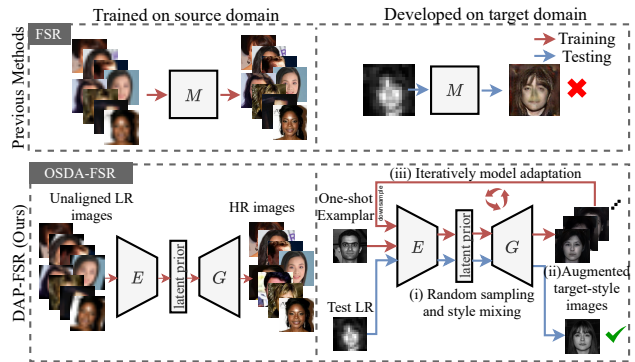

Figure 1. Conventional FSR methods achieve good performance on the source dataset, but are prone to fail on the target dataset due to the domain gap. Our proposed method effectively adapts the model by leveraging only one-shot example.

has achieved great success in recent years [51, 53, 52, 48, 54, 49, 56, 5, 8, 61].

Previous FSR methods usually presume training and testing LR faces are captured from the same domain. When testing LR faces resemble the training ones, previous works achieve authentic upsampled HR faces. However, in practice, the domain gap between testing images and training ones is inevitable due to different imaging equipment, illumination conditions, *etc*. As shown in the upper right of Figure 1, previous state-of-the-art FSR methods fail to upsample HR authentically due to the large domain gap between the target domain (testing) and source domain (training). Considering FSR models would be deployed in different scenarios, it is very inefficient to re-train every deployed FSR model by collecting large-scale data from the corresponding target domain. Therefore, only using a few samples, ideally one example, to efficiently update an FSR model is highly desirable.

In this paper, we aim to super-resolve LR faces that exhibit an obvious domain gap by only leveraging one-shot exemplar from the target domain. We name this task as One-Shot Domain Adaption for Face Super-Resolution (OSDA-FSR). Different from conventional FSR methods [51, 8, 37], two challenges are naturally raised: (i) how

---

to design a FSR network architecture that is intrinsically suitable for efficient adaptation; and (ii) how to explore one example to bridge the domain gap since simply fine-tuning an FSR network with one example is ineffective.

To address these challenges, we present a novel Domain-Aware Pyramid based Face Super-Resolution network, namely DAP-FSR network. Our DAP-FSR contains two parts: a domain-aware pyramid encoder and an upsampling decoder. Our DAP-FSR encoder is designed to extract the latent representations by leveraging the multi-scale features from the input LR faces. Considering LR faces may be unaligned, we propose an Instance Spatial Transformer Networks (ISTN) to align LR faces inspired by [21]. In this way, we facilitate the latent representation learning and face upsampling processes by aligning the LR faces into the canonical view. Motivated by the powerful architecture of StyleGAN [23, 24], an image generation network, we construct our upsampling decoder. Once we obtain the latent representations, we feed those representations to our DAP-FSR decoder to hallucinate high-quality HR face images.

To tackle the problem of super-resolving LR faces in a new domain without the need for tremendous data collection, we propose a Domain-Aware latent Mixing and Model Adaptation algorithm (DAMMA). In a nutshell, our DAMMA algorithm is able to adapt the model trained on the source domain to the target domain by exploring only the one-shot example. As illustrated in Figure 1, when a target domain example is given, DAP-FSR network first extracts its latent representations. Then, supervised by the given one-shot example, we learn a soft mixture weight to mix the target latent representations with random-sampled source latent ones. In this fashion, the newly generated faces will resemble the target domain faces and we significantly augment the target-style data. By constrained fine-tuning our decoder with the augmented images, our network is gradually adapted from the source domain to the target domain. After iteratively updating the soft mixing weight and adapting our decoder, our DAP-FSR attains authentic target domain HR faces.

Our main contributions are summarized as follows,

- We propose a novel domain-aware pyramid-based face super-resolution network, named DAP-FSR network, to efficiently upsample cross-domain LR face images by peeking at one-shot target domain example.
- We present a simple yet effective domain-aware latent mixing and model adaptation algorithm (DAMMA) to adapt our DAP-FSR to the target domain. Our DAMMA generates target-style alike faces to adapt the upsampling decoder in DAP-FSR by fully exploiting the one-shot example.
- To the best of our knowledge, our method is the first attempt to super-resolve cross-domain LR face images, making our method more practical.
- Our proposed DAP-FSR can be adapted to a target domain effectively and is also robust to unaligned LR faces. Experiments on three constructed cross-domain face super-resolution benchmarks validate the superior performance of our proposed approach compared to the state-of-the-art methods.

2. Related Work

Face super-resolution. Face super-resolution (FSR), also known as face hallucination, aims at establishing the intensity relationships between input LR and output HR face images from the same domain.

Traditional holistic appearance-based methods firstly leverage a parameterized model to represent faces and then construct the mappings between LR and HR faces. Some representative models super-resolve HR faces from LR ones by adopting global linear mapping [45, 29], or optimal transport [25]. However, they require input LR images to be aligned to a canonical pose and HR faces in the database to share similar facial expressions. Later on, part-based approaches have been proposed to relax the strict requirements in holistic appearance-based methods. Part-based face hallucination algorithms [36, 43, 47] firstly extract local facial regions and then upsample them separately.

Taking advantage of the powerful feature representation of deep neural networks, deep learning based face super-resolution methods [65, 51, 52, 50, 59, 65, 18, 8, 37, 60] have been proposed and achieved promising results. Several methods exploit prior knowledge, such as facial attributes [49], parsing maps [8], facial landmarks [48, 7, 6] and identity [57, 42], to advance the upsampling performance. However, when LR faces are captured from another domain, such as different imaging conditions, existing methods may fail to super-resolve them photo-realistically. Moreover, when the new domain data is not abundantly available, it would be difficult to retrain FSR networks with such a limited number of samples. In this paper, we make the first attempt to address this challenging scenario in a data efficient manner.

One-shot Domain Adaptation To overcome the need of large-scale training data and improve the adaption ability of models on new domains, many works have been extensively proposed [13, 30, 38, 4, 3, 44, 33, 10, 12, 11]. Early one/few-shot based classification tasks [9] construct generative models from shared appearance priors across classes for classification. Recently, a new stream of works focuses on using meta-learning to quickly adapt models to novel tasks [13, 39, 41]. However, these one/few-shot methods are mainly applied to different classification tasks without taking domain gaps between image-pairs into account.

Pix2Pix [20] and CycleGAN [64] have been proposed as image-to-image translation networks. However, due to the scarcity of samples in the target domain, these methods might not be suitable for transferring from the source domain to the target one with few samples. To mitigate the data hungry problem of deep neural networks, several works employ shared [30] or partially shared [26] latent space assumption to conduct image-to-image translation tasks, such as style transfer [19, 30] and face generation [46]. Since these methods only address the domain gap without learning the mapping between LR and HR images, they are not suitable for face hallucination.
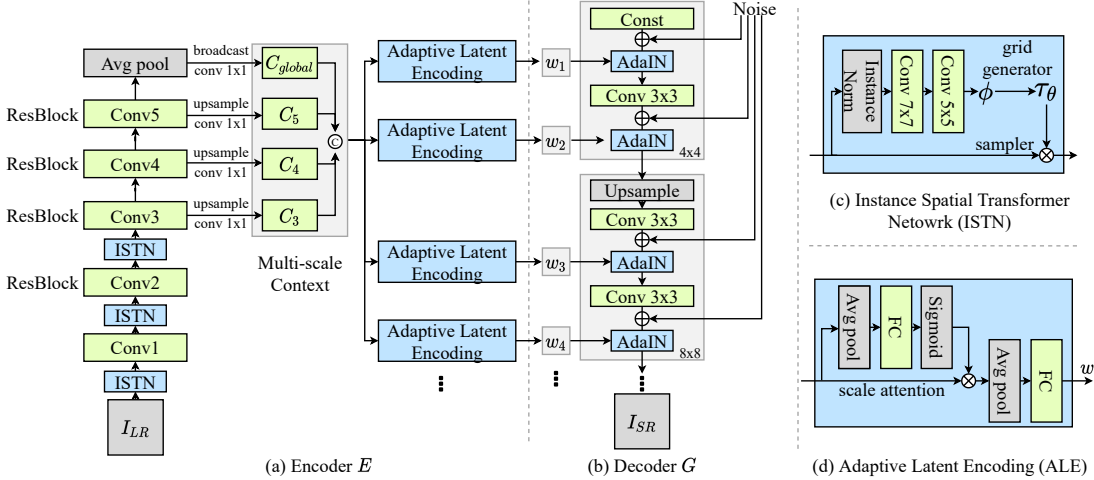
Figure 2. Illustration of our DAP-FSR architecture. (a) The encoder network. Feature maps from different spatial resolution are up-sampled and concatenated as the multi-scale pyramid context. Each Adaptive Latent Encoding (ALE) module dynamically attends the multi-scale context to generate the latent representation $\boldsymbol{w}_i$. (b) The decoder network, where the HR images are generated based on the latent representations. (c) The Instance Spatial Transformer Network (ISTN) learns the style-invariant affine transformation matrix to adjust the unaligned LR images. (d) The detailed Adaptive Latent Encoding module, where the channel-wise feature attention is learned to adaptively capture the multi-scale information of the input images.

3. Task Definition: One-shot based FSR

Conventional Face Super-Resolution (FSR) methods aim to learn a face super-resolution model M that generates a high-resolution super-resolved face image $I_{SR} \in \mathbb{R}^{H \times W}$ from a low-resolution one $I_{LR} \in \mathbb{R}^{h \times w}$, as follows:

$$I_{SR} = M\left(I_{LR}\right). \tag{1}$$

The goal of the FSR task is to make the reconstructed image I_{SR} best recover its corresponding high-resolution version I_{HR}. In the conventional face hallucination setting [29, 36, 51], an FSR model M is trained and evaluated on the $\{(I_{LR}, I_{HR})\}$ pairs from the same source domain. However, as illustrated in Figure 1, when LR images come from another target domain, a pre-trained model M might fail to generalize well to the new domain data and the quality of super-resolved HR images will degrade severely.

Inspired by previous domain adaptation works [40], we formulate our task as One-Shot Domain Adaptation for Face Super-Resolution (OSDA-FSR). In general, OSDA-FSR can be divided into two stages, *i.e.*, a procurement stage and a deployment stage, based on the real-world application scenario. In the procurement stage, an FSR model is trained on the large-scale source dataset with N_s HR and LR image pairs, denoted as $\mathcal{D}_s = \{(I_{LR}^s, I_{HR}^s)_i\}_{i=1}^{N_s}$. In this stage, image reconstruction objectives will be employed to optimize the model parameters. However, in the deployment stage, a trained model might encounter an unknown data distribution shift in a target domain. In this case, a deep model may fail to super-resolve LR faces in a target domain without knowing any information about the new domain.

Although collecting data and re-training a network can solve this issue, it might be inefficient and time-consuming when deploying deep models in many different real-world scenarios. Therefore, we aim to use only a few examples, *e.g.*, K LR-HR pairs $\mathcal{D}_t = \{(I_{LR}^t, I_{HR}^t)_i\}_{i=1}^{K}$, to effectively adapt the pre-trained model M. Without the loss of generality, we focus on the most challenging case where $K = 1$. In other words, we will exploit the one-shot exemplar to minimize the domain gap and then hallucinate the target domain LR faces.

4. Proposed Method

Overview. The general goal of OSDA-FSR task is to transfer the model from the trained source domain to the target domain by fully exploiting the given one-shot example. To achieve this goal, the key idea of our approach is to adapt the model towards the target domain by enriching the target-style samples beyond the solely given one-shot exemplar. We present a Domain-Aware Pyramid-based Face Super-Resolution (DAP-FSR) network to super-resolve input LR images to output HR images, as shown in Figure 2. Our DAP-FSR firstly obtains the semantic latent representations from an unaligned LR face image by the encoder network and then generates the high-quality HR images from these latent representations by the upsampling decoder network.

Given an LR image in the target domain, our DAP-FSR network first extracts the latent representations. However, due to the existing large domain gap, the latent representations of target domain LR images may not lie on the manifold of the source domain ones, thus causing inferior upsampled results. To address this problem, we propose to project the latent representations of the target one-shot exemplar to the closest one in the source domain. We then synthesize random images sharing similar styles with the target domain by mixing randomly sampled source and the extracted target domain latent representations. These generated samples will be in turn used to optimize our upsam-

pling network. In this fashion, the latent representation manifold will gradually shift to the target domain and we can super-resolve target domain LR images even with only one exemplar.

4.1. Domain Aware Pyramid-based FSR

Choice of decoder and latent space. Due to the advanced network architecture, StyleGAN [23, 24] obtains phenomenal high-resolution and photo-realistic images. Recent work [37] also demonstrates the possibility that employing a pre-trained StyleGAN, HR faces can be found from the given LR inputs. More importantly, decouple the training of an encoder and a decoder would allow us to achieve larger upscaling factors while being less restricted by GPU memory. Therefore, we choose the StyleGAN architecture as the upsampling decoder in our DAP-FSR.

Former work [37] demonstrates that the multi-layer disentangled latent space $\mathcal{W}+$ in StyleGAN is more representative to depict an image than the normalized Gaussian distribution space $\mathcal{Z}$. Furthermore, the layer-wise corresponding AdaIN modules in StyleGAN can also facilitate us to transfer domain-specific characteristics when we adapt our trained upsampling decoder to a target domain. Hence, to fully utilize the power of StyleGAN, we adopt the $\mathbf{w} \in \mathbb{R}^{l \times d_w}$ as our latent representations to better encode the LR images, where l is the layer number and d_w is the latent representation dimension.

Latent representation learning. Unlike PLUSE [37] that optimizes a latent representation $\mathbf{w} \in \mathcal{W}+$ by minimizing the pixel-wise reconstruction loss between the downsampled version of upsampled HR image and the input image, we introduce an encoder to extract latent representations of the input LR faces. Doing so allows us to address unaligned LR faces and handle the domain gap by fine-tuning our upsampling decoder, while PLUSE cannot handle the domain gap and face misalignments as its decoder (*i.e.*, pre-trained StyleGAN) is fixed and only $\mathbf{w}$ is updated during iterations.

Recall that in the StyleGAN, each latent representation controls a certain level of image details. Hence, our encoder aims to adaptively predict latent representations from an enhanced multi-scale context feature. Toward this goal, we develop an adaptive latent encoding (ALE) module that is able to generate latent representations for the upsampling decoder at different scales adaptively.

Here, we employ ResNet50 as our encoder to extract multi-scale feature maps at the `conv3`, `conv4`, `conv5` and average pooling layers, denoted as $C_3, C_4, C_5, C_{global}$, as shown in Figure 2. Then, each ALE generates multiscale latent representations $\mathbf{w}_i$ for the decoder by attending the multi-scale features adaptively. Then, the latent representations are fed to our upsampling decoder for face hallucination.

Robust against unaligned LR faces. Previous face hallucination methods [32, 36, 37] often assume LR faces are precisely aligned beforehand. However, such an assumption hardly holds in real application scenarios. Inspired by the works [53, 61], we estimate the transformation of LR images and warp them to the canonical position by the spatial

---

Algorithm 1: Domain-Aware Latent Mixing and Model Adaptation

Input: Initialized DAP-FSR model $M = (E, G)$ trained on source dataset $\mathcal{D}_s$, one-shot exemplar $\{I_{LR}^t, I_{HR}^t\} \in \mathcal{D}_t$, initialized latent code mixing weight $\boldsymbol{\alpha}_0$, AdaIN parameter ϕ in G ,learning rate ξ, η

Output: Adapted model M_{ϕ^*}

1 while *do not converge* **do**

2   Generate $\mathbf{w}^t$ by manifold preserving projection as Eq. (2);

3   Sample a batch of source latent codes: $\mathbf{w}_s = \mu_{\mathbf{w}} + \sigma_{\mathbf{w}}\epsilon, \epsilon \sim \mathcal{N}(0, 1)$;

4   Initialize latent code mixing weight: $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}_0$;

5   **for** *i=1,2,3,...,n* **do**

6   Update mixing weight by Eq. (7): $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \xi\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha})$;

7   Generate mixing latent codes $\mathbf{w}^m$ by Eq. (6);

8   Update model parameters by Eq. (5): $\phi \leftarrow \phi - \eta\nabla_{\phi}\mathcal{L}(\phi)$;

9   **end**

10 end

11 Return final model weight ϕ as ϕ^*;

---

transformation network (STN) [21]. Therefore, our network is robust against unaligned LR faces with in-plane rotations, translations and scale changes. The detailed architecture of spatial transformation layers are illustrated in Figure 2(c).

More importantly, unlike previous FSR models [53, 61] that use STNs, we apply an instance normalization layer to the feature maps before computing the transformation parameters in our instance spatial transformer network (ISTN) module. This allows us to obtain style-invariant feature maps. Therefore, even when target-domain LR faces are provided, our ISTN layers are still able to align them to the up-right position, potentially facilitating the following domain adaptation process. Thus, our decoder can focus on super-resolving high-quality HR faces while preserving the latent representations from being affected by misaligned input LR faces.

Manifold preserving encoding. Previous work [63] shows that it is possible to invert an arbitrary image, even not a face image, into style latent space $\mathcal{W}+$. However, such deduced latent codes are not aligned with the semantic knowledge prior learned by $G(\cdot)$ and lose the versatile image editing capability. In our OSDA-FSR task, the situation will become even worse where a domain gap between the source and target domains exists.

To overcome these drawbacks, we explicitly constrain the output of the encoder $E(\cdot)$ in the feature space of G. Particularly, instead of directly predicting the style latent codes, we predict the offset scale *w.r.t.* the mean $\mu_{\mathbf{w}}$ and variance $\sigma_{\mathbf{w}}$ of the latent representations of G. To be specific, our DAP-FSR model maps the encoded representations of LR images to the latent representations $\mathbf{w}$ of the

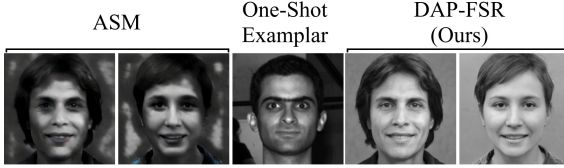| ASM	One-Shot Exemplar	DAP-FSR (Ours)

Figure 3. Compared to the style-transfer based method ASM [35] (left), given only one-shot target domain exemplar (ExtendedYaleB), our method (right) efficiently generates authentic target-style images from the source domain (CelebA).

decoder, as follows:

$$\mathbf{w} = \mu_\mathbf{w} + E(I_{LR})\sigma_\mathbf{w}, \qquad (2)$$

where $\mu_\mathbf{w}$ and $\sigma_\mathbf{w}$ are fixed during the encoder training process. Therefore, using Eq. (2), we can explicitly constrain the latent representation output by our encoder to lie in the latent representation space $\mathcal{W}+$ of our decoder G.

Network optimization. Our encoder E is trained using two losses. We employ the pixel-wise reconstruction loss $\mathcal{L}_{mse}$ to enforce reconstructed HR images to be close to their HR ground-truth I_{HR},

$$\mathcal{L}_{mse} = \|I_{HR} - G(\mathbf{w}))\|_2. \qquad (3)$$

In addition, we also introduce the perceptual loss to enforce the feature-wise similarity,

$$\mathcal{L}_{percept} = \|F(I_{HR}) - F(G(\mathbf{w}))\|_2, \qquad (4)$$

where F denotes the perceptual feature extractor. In our experiments, we extract features from `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1` layers in VGG-19 with equal weights. In our final objective, we also treat the image intensity similarity and feature similarity equally, and the objective is defined as,

$$\mathcal{L}(\theta) = \mathcal{L}_{mse} + \mathcal{L}_{percept}, \qquad (5)$$

where θ is the trainable parameters of our network. Note that our upsampling decoder and encoder are trained individually and thus our decoder is fixed during training our encoder.

4.2. Peeking at One-Shot Exemplar

Towards target-domain image generation. Benefiting from the encoder design in our DAP-FSR network, we can encode the given one-shot target domain HR image I^t_{HR} into a latent representation $\mathbf{w}^t$. However, using only one-shot exemplar does not suffice to transfer our decoder to the target domain, and will lead to an over-fitting problem. As explained in [23], the latent codes of the StyleGAN control the coarse, medium, fine attributes of generated images at different style layers. Thus, we also regard the latent code $\mathbf{w}^t$ as an interpretable representation of a target domain face. Moreover, we can generate a large number of domain-specific (i.e., style-consistent) face images with I^t_{HR}. Specifically, for a latent code $\mathbf{w}^s$ randomly sampled from the latent representation manifold of the source domain, we mix it with $\mathbf{w}^t$ in a layer-wise manner so that

a generated image I^m inherits the the target domain style from I^t. The mixing procedure is defined as:

$$\mathbf{w}^m_i = (1 - \boldsymbol{\alpha}_i)\mathbf{w}^t_i + \boldsymbol{\alpha}_i\mathbf{w}^s_i, \qquad (6)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^l$ is a layer-wise soft weight for mixing latent representations. In this manner, we effectively enlarge the number of target domain examples from the given one-shot exemplar by $G(\mathbf{w}^m)$. In Figure 3, compared with a style transfer based method (i.e., ASM [35]), our method is able to generate more natural style-consistent images while preserving the identity.

Learning soft mixing weight. When mixing the latent representations of random sampled $\mathbf{w}^s$ and the target sample $\mathbf{w}^t$, we preserve the image content information by applying a feature-wise intensity consistent loss $\mathcal{L}_c$ and enforce the domain information to be transferred by employing a style similarity loss $\mathcal{L}_s$. Here, we learn a soft weight $\boldsymbol{\alpha}$ to mix the latent codes of the source and target domain instead of manually selecting a certain layer, and the optimization process is formulated as,

$$\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}_c + \mathcal{L}_s, \qquad (7)$$

$$\mathcal{L}_c = \|F(G(\mathbf{w}^m)) - F(G(\mathbf{w}^s))\|_2, \qquad (8)$$

$$\mathcal{L}_s = \|\mu(F(I^t) - \mu(F(G(\mathbf{w}^m)))\|_2 + \\ \|\sigma(F(I^t)) - \sigma(F(G(\mathbf{w}^m)))\|_2. \qquad (9)$$

where μ and σ denote the mean and variance of the extracted features respectively, and F is the same perceptual extractor in Eq. (4).

Model updating by constrained adaptation. After we generate a batch of random images exhibiting the same target domain style, our next step is to adapt our model towards the target domain. The most straightforward way is to fine-tune the entire decoder G directly on our generated target-domain alike samples. However, when the number of training examples is limited, especially in our case, fine-tuning the whole network weights often leads to over-fitting and may potentially destroy the learned knowledge prior in G. Instead of fine-tuning the entire decoder weights, we constrain the fine-tuning on a subset of the decoder parameters. To be specific, we only adapt the affine transform parameters in the AdaIN module. By restricting the trainable parameters, our model can be effectively adapted to the target domain while preserving the semantic knowledge, i.e., natural face structure. The overall pipeline of our algorithm is illustrated in Algorithm 1.

4.3. Training and Inference

Our training process consists of two main stages, the procurement stage and development stage. In the procurement stage, we first train our decoder G following the protocols of StyleGAN and then only train the encoder model E on the source dataset by Eq. (5) while fixing the parameters of G. After training, our DAP-FSR is able to super-resolve HR faces from LR faces with an upscaling factor up to $\times 64$. In the development stage, we peek at the one-shot exemplar from the target domain and adapt our model to the target

Table 1. Comparison with state-of-the-art methods. Results are reported on three benchmarks noted as **source → target**. 'Source only' denotes the methods only using source dataset for training, while 'one-shot' denotes the methods exploring one-shot exemplar on the target dataset. ↑ indicates that higher is better, and ↓ that lower is better.

	Method	CelebA → ExtendedYaleB				CelebA → MultiPIE				MultiPIE → ExtendedYaleB			
		LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑
source only	Bicubic	0.52	0.31	19.94	0.46	0.55	0.27	17.11	0.39	0.54	0.31	17.70	0.43
	PUSLE [37][CVPR'20]	0.40	0.38	20.18	0.46	0.46	0.36	14.63	0.37	0.42	0.27	17.02	0.46
	MTDN [55][IJCV'20]	0.39	0.32	17.74	0.45	**0.38**	0.38	18.00	0.52	0.47	0.20	18.67	0.43
	CPGAN [61][CVPR'20]	0.40	0.28	17.03	0.47	0.40	0.31	18.61	0.52	0.45	0.24	18.80	0.44
	DAP-FSR (Ours)	**0.38**	**0.41**	**20.39**	**0.49**	**0.38**	**0.40**	**19.15**	**0.54**	**0.41**	**0.34**	**19.28**	**0.46**
one-shot	PULSE+ASM [35][NeurIPS'20]	0.44	0.32	20.47	0.47	0.49	0.32	17.87	0.41	0.44	0.23	17.89	0.43
	MTDN+ASM	0.42	0.27	19.01	0.48	0.44	0.33	19.38	0.53	0.52	0.25	19.11	0.47
	CPGAN+ASM	0.49	0.26	18.42	0.42	0.49	0.29	19.29	0.55	0.51	0.23	19.19	0.49
	DAP-FSR (Ours)	**0.36**	**0.46**	**22.32**	**0.55**	**0.36**	**0.44**	**21.00**	**0.61**	**0.39**	**0.40**	**20.43**	**0.51**

domain by employing our proposed Algorithm 1. During inference, we test our adapted model on the whole target dataset and report the super-resolution performance. Note that, we only see one-shot image from the target domain and all other testing images are *never seen* during training.

5. Experiments

In this section, we conduct extensive experiments to evaluate our DAP-FSR framework. Since we focus on the OSDA-FSR task, we mainly compare with the state-of-the-art in this scenario.

5.1. Datasets and Evaluation Protocols

Benchmarks. Current FSR benchmarks conduct training and testing within the same domain, and do not support the setting of the cross-domain OSDA-FSR task. Therefore, We propose three benchmarks to evaluate the performance of our DAP-FSR, *i.e.*, CelebA [34] → Multi-PIE [16], CelebA → ExtendedYaleB [14], and Multi-PIE → ExtendedYaleB. In particular, CelebA dataset contains large-scale in-the-wild face images, Multi-PIE and ExtendedYaleB datasets comprise indoor face images captured in different poses and illumination conditions. We select 10 different illumination and pose condition data splits in Multi-PIE and ExtendedYaleB, respectively. The adaptation performance is evaluated with a given exemplar in each split and then the final reported performance is averaged over all the splits.

Evaluation metrics. We report the quantitative results using the average Peak Single-to-Noise Ratio (PSNR), Structural SIMilarity scores (SSIM) following the common FSR practice [53, 61]. Furthermore, we also employ the Learned Perceptual Image Patch Similarity (LPIPS) [58] and Face Image Quality (FIQ) [17] to evaluate the quality and authenticity of super-resolved faces. The PSNR, SSIM, LPIPS metrics are calculated between the reconstructed HR images I_{SR} and the ground-truth HR images I_{HR}. The FIQ is a non-reference metric for face quality assessment, which is calculated only on I_{SR}.

5.2. Implementation Details

In our experiments, we crop the aligned faces and resize them to 128×128 pixels to achieve ground-truth HR images. In real-world applications, we do not assume that the input LR faces are perfectly aligned. Following [55],

we apply affine transformations, including rotations, translations and scaling, to HR faces and then downsample them to 16×16 pixels as our LR face images. We use the author-provided codes of PULSE [37], MTDN [55] and CP-GAN [61]. For comparison fairness, we adopt the same training protocols for all the methods. To alleviate the influence of the selected one-shot exemplar, we run the proposed method for ten times with different randomly selected one-shot exemplars in each task and report the averaged results.

5.3. Comparisons with the State-of-the-Art

Qualitative comparisons. We first conduct qualitative comparisons with the state-of-the-art methods on three OSDA-FSR benchmarks in Figure 4.

CPGAN [61] and MTDN [55] can super-resolve LR images well and deal with unaligned LR input faces successfully in the source domain. However, these methods do not take the domain gap into account, and lack an efficient mechanism to address LR images from a new domain. Therefore, their final reconstructed HR images from target domain LR faces suffer from severe artifacts. Although collecting a large number of target domain data and then retraining the networks can solve the above issue, doing so is time-consuming and does not provide a data-efficient solution to OSDA-FSR.

PULSE [37] traverses the high-resolution face image manifold and searches images whose downsampled versions are close to the given LR images. Although realistic images are achieved, this method requires input LR images to be perfectly pre-aligned. When LR images are unaligned, the reconstructed HR images are enforced to match the intensities of LR faces. This will lead to severe changes of face identities, as seen in Figure 4. Moreover, PULSE does not consider the domain gap. Due to the data distribution shift between the source and target domains, PULSE fails to super-resolve HR faces sharing the same style as the target domain images.

In contrast, as seen in Figure 4, our method achieves superior performance compared to the other competing methods. Although input LR images are unaligned, our DAP-FSR still produces visually appealing HR faces which are close to their HR ground-truth. Notably, our upsampled faces also exhibit style-consistency with respect to the given one-shot target domain exemplar. This demonstrates the transfer ability of our method. Note that our method is

Figure 4. Comparisons with state-of-the-art methods on CelebA→ExtendedYaleB, CelebA→MultiPIE and MultiPIE→ExtendedYaleB benchmarks under the OSDA-FSR setting. Our method achieves high-quality, style-consistent HR faces and is also robust against unaligned LR inputs.

actually able to super-resolve LR faces with an upscaling factor up to $64\times$, and for fair comparisons with the state-of-the-art methods, we only show HR faces in the same resolution as other methods. To the best of our knowledge, our DAP-FSR network is *the first attempt to super-resolve cross-domain LR images with only one target-domain exemplar*, and achieves superior super-resolution results.

To further validate the generalization ability, in Figure 5, we show the FSR results of tiny faces *in-the-wild* [2] under *real-world unconstrained conditions*, where the ground-truth HRs are unavailable. Here, LR faces may undergo different poses, blurs, noises, etc. All the models are trained on the CelebA source dataset and adapted to the target domain using the given one-shot HR example. Moreover, in Figure 6, we also conduct cross-domain FSR experiments on near infrared (NIR) face images [28] as a target domain. Our DAP-FSR still outperforms the other competing methods, demonstrating the generalization ability of our method.

Quantitative comparisons. As indicated in Table 1, we report the LPIPS, FIQ, PSNR and SSIM metrics on three OSDA-FSR benchmarks, respectively. Our proposed DAP-FSR outperforms the state-of-the-art methods significantly, especially on the perceptually-driven metrics, *i.e.*, LPIPS and FIQ. This indicates that our super-resolved target domain HR faces not only resemble their ground-truth but also are photo-realistic. More importantly, our DAP-FSR consistently performs better than other methods on all the benchmarks. Thanks to our dedicated network design, we are able to align and upsample target domain LR faces, simultaneously. In particular, DAP-FSR reconstructs high-quality face images and outperforms the second best method PULSE on unaligned images by a margin of $+43\%$ ($0.32 \rightarrow 0.46$) in FIQ on the benchmark CelebA→ExtendedYaleB.

To address the domain gap, a straightforward idea is fine-tuning the source-trained FSR model with the augmented
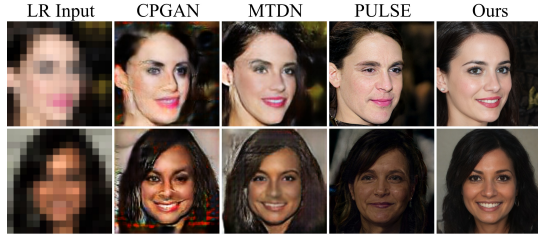
Figure 5. Comparisons with state-of-the-art methods on tiny faces in-the-wild [2] under real-world unconstrained conditions.
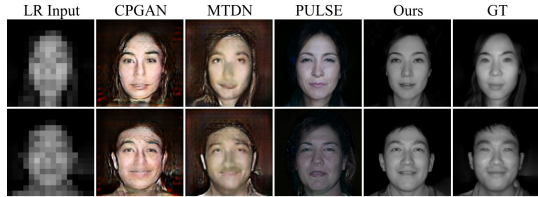

Figure 6. Comparisons with state-of-the-art methods on near-infrared (NIR) sensor captured faces [28].

Table 2. Ablations on different configurations of the network architecture (A,B,C) and different configurations of the adaptation algorithm (D,E,F). ↑ indicates the higher the better, and ↓ indicates the lower the better.

| Configuration | CelebA → ExtendedYaleB | | | |
	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑
Baseline network	0.48	0.28	17.64	0.44
A + Multi-scale features	0.46	0.30	17.82	0.44
B + Multi-STN modules	0.43	0.34	18.41	0.45
C + Predict offset scale	0.38	0.41	20.39	0.49
D + Style mixing examples	0.38	0.41	21.97	0.52
E + Soft mixing weight	0.38	0.42	22.10	0.54
F + Constrained adaptation	0.36	0.46	22.32	0.55

Table 3. Comparisons on one-shot adaptation augmentation strategies. ↑ indicates the higher the better, and ↓ the lower the better.

| Methods | CelebA → ExtendedYaleB | | | |
	LPIPS ↓	FIQ ↑	PSNR ↑	SSIM ↑
Direct fine-tuning	0.44	0.30	20.11	0.45
Style Transfer [19]	0.42	0.37	20.16	0.46
ASM [35]	0.40	0.38	20.71	0.50
DAP-FSR (Ours)	0.36	0.46	22.32	0.55

target samples. Thus, we employ a style-transfer-based method ASM [35] to augment new training samples from the one-shot target domain exemplar, and then fine-tune the FSR models. We name these the combination as +ASM in Table 1. As indicated by Table 1, applying style transfer cannot fully establish the facial detail correspondences between the source and target domains, thus leading to performance degradation.

Furthermore, benefiting from our designed one-shot adaptation algorithm, we transfer our network to the target domain effectively. Therefore, our quantitative results are better than the results of MTDN+ASM and PULSE+ASM. Owing to our encoder-decoder design, our method is also more efficient and effective compared to the decoder-only based method PULSE. After training, our DAP-FSR hallucinates LR faces in a feed-forward manner and runs ×150 faster than PULSE, which provides a high application potential in the real-world scenario.

5.4. Ablation Analysis

In our ablation analysis, we conduct all the experiments on the CelebA→ExtendedYaleB benchmark.
Effectiveness of network design. We analyze the effect of each component in our network design in Table 2. Compared to a straightforward approach that predicts the latent representations at the end of the backbone, our network adaptively explores the abundant multi-scale features (Config A). It is a long-standing shortcoming that CNN is sensitive to rotations. Our multiple ISTN design effectively handle this problem (Config B), thus being robust against unaligned LR images. We also illustrate that it is vital to explicitly constrain the predicted latent representations on the manifold (Config C).
Effectiveness of one-shot domain adaptation. Table 2 indicates the impact of each component in Algorithm 1 on the OSDA-FSR performance. In our method, we effectively enrich the training samples by mixing the latent representa-

tions between the source and target domain faces (Config D). Compared to the configuration without exploring the one-shot exemplar (Config C), we observe that Config D achieves better super-resolution performance. This implies our method fully exploits the one-shot target exemplar to bridge the domain gap.

By applying the soft mixing weight (Config E), we further improve the super-resolution performance. This indicates that our soft mixing strategy is more effective than simply replacing the last three final layers of the latent representations between the source and target domain images as done in Config D. As fine-tuning the whole decoder network may lead to over-fitting and destroy the learned face priors, we constrain the optimization space and only modify the AdaIN parameters to improve performance (Config F).

We also compare with other target domain augmentation methods, including Style Transfer and ASM. Specifically, these are employed to enlarge the target domain examples and then we constrained fine-tune our model using the augmented data. As indicated in Table 3, our method significantly facilitates the model adapting to the target domain, thus achieve better super-resolution performance.

6. Conclusion

In this paper, we addressed a more challenging and practical face super-resolution task, where a domain gap between the training and testing data exists. To tackle this problem, we proposed a new Domain-Aware Pyramid-based Face Super-Resolution network (DAP-FSR) that is able to super-resolve unaligned low-resolution ones from a target domain effectively by leveraging only one target domain exemplar. Our approach bridges the domain gap by fully exploiting the given exemplar from the target domain as well as our designed soft mixing strategy which significantly enlarges the number of the training samples. Extensive experiments demonstrate our method is able to super-resolve cross-domain LR faces and outperforms the state-of-the-art methods significantly. We hope that our work will also motivate future research on the low-shot FSR task.

References

[1] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–30, 2018. 7, 8

[3] Sagie Benaim and Lior Wolf. One-shot unsupervised cross domain translation. In *Advances in Neural Information Processing Systems*, pages 2104–2114, 2018. 2

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 2

[5] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2018. 1

[6] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[7] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 185–200, 2018. 2

[8] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2018. 1, 2

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 2

[10] Qianyu Feng, Guoliang Kang, Hehe Fan, and Yi Yang. Attract or distract: Exploit the margin of open set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7990–7999, 2019. 2

[11] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, Mingliang Xu, and Yi Yang. Cascaded revision network for novel object captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3413–3421, 2020. 2

[12] Qianyu Feng, Zongxin Yang, Peike Li, Yunchao Wei, and Yi Yang. Dual embedding learning for video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2

[14] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001. 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 1

[16] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–8, 2008. 6

[17] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 6

[18] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1689–1697, 2017. 2

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2, 8

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 2

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems(NIPS)*, pages 2017–2025, 2015. 2, 4

[22] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 4, 5

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 4

[25] Soheil Kolouri and Gustavo K Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4876–4884, 2015. 2

[26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[27] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1911–1920, 2021. 1

[28] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013. 7, 8

[29] Ce Liu, Heung-Yeung Shum, and William T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007. 2, 3

[30] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 2

[31] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 1

[32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[33] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*, 2019. 2

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, page 3730–3738, December 2015. 6

[35] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 5, 6, 8

[36] Xiang Ma, Junping Zhang, and Chun Qi. Hallucinating face by position-patch. *Pattern Recognition*, 43(6):2224–2236, 2010. 2, 3, 4

[37] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020. 1, 2, 4, 6

[38] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 6670–6680, 2017. 2

[39] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[40] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5880–5888, 2019. 3

[41] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 2

[42] Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, and Piotr Koniusz. Identity-preserving face recovery from stylized portraits. *International Journal of Computer Vision*, 127(6-7):863–883, 2019. 2

[43] Marshall F Tappen and Ce Liu. A bayesian approach to alignment-based image hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 236–249, 2012. 2

[44] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*, 2021. 2

[45] Xiaogang Wang and Xiaoou Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems,*

Man, and Cybernetics, Part C (Applications and Reviews), 35(3):425–434, 2005. 2

[46] Chao Yang and Ser-Nam Lim. One-shot domain adaptation for face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5921–5930, 2020. 2

[47] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang. Hallucinating compressed face images. *International Journal of Computer Vision*, 126(6):597–614, 2018. 2

[48] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 217–233, 2018. 1, 2

[49] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 908–917, 2018. 1, 2

[50] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[51] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 318–333, 2016. 1, 2, 3

[52] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, page 4327–4333, 2017. 1, 2

[53] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3760–3768, 2017. 1, 4, 6

[54] Xin Yu and Fatih Porikli. Imagining the unimaginable faces by deconvolutional networks. *IEEE Transactions on Image Processing*, 27(6):2747–2761, 2018. 1

[55] Xin Yu, Fatih Porikli, Basura Fernando, and Richard Hartley. Hallucinating unaligned face images by multiscale transformative discriminative networks. *International Journal of Computer Vision*, 128(2):500–526, 2020. 6

[56] Xin Yu, Fatemeh Shiri, Bernard Ghanem, and Fatih Porikli. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2148–2164, 2019. 1

[57] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European conference on computer vision (ECCV)*, pages 183–198, 2018. 2

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[59] Yang Zhang, Ivor Tsang, Yawei Luo, Changhui Hu, Xiaobo Lu, and Xin Yu. Recursive copy and paste gan: Face hallucination from shaded thumbnails. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[60] Yang Zhang, Ivor W Tsang, Jun Li, Ping Liu, Xiaobo Lu, and Xin Yu. Face hallucination with finishing touches. *IEEE Transactions on Image Processing*, 30:1728–1743, 2021. 2

[61] Yang Zhang, Ivor W Tsang, Yawei Luo, Chang-Hui Hu, Xiaobo Lu, and Xin Yu. Copy and paste gan: Face hallucination from shaded thumbnails. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7355–7364, 2020. 1, 4, 6

[62] Wang Zhibo, Yu Xin, Lu Ming, Wang Quan, Qian Chen, and Xu Feng. Single image based portrait relighting via explicit multiple channel modeling. In *Siggraph Asia*, pages 1–13, 2020. 1

[63] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Eur. Conf. Comput. Vis.*, 2020. 4

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2223–2232, 2017. 2

[65] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 614–630, 2016. 2