HIME: Efficient Headshot Image Super-Resolution with Multiple Exemplars

Xiaoyu Xiang[1,2*], Jon Morton[2], Fitsum A. Reda[2†], Lucas D. Young[2†], Federico Perazzi[2†],
Rakesh Ranjan[2], Amit Kumar[2], Andrea Colaco[2†], Jan P. Allebach[1]
[1]Purdue University, [2]Meta Reality Labs
{xiang43,allebach}@purdue.edu, {jamorton,rakeshr,akumar14}@meta.com,
{fitsum.reda,lucasyoung482}@gmail.com, fdp@bendingspoons.com, andrea@andreacolaco.info

Abstract

*A promising direction for recovering the lost information in low-resolution headshot images is utilizing a set of high-resolution exemplars from the same identity. Complementary images in the reference set can improve the generated headshot quality across many different views and poses. However, it is challenging to make the best use of multiple exemplars: the quality and alignment of each exemplar cannot be guaranteed. Using low-quality and mismatched images as references will impair the output results. To overcome these issues, we propose the **H**eadshot **I**mage Super-Resolution with **M**ultiple **E**xemplars network (HIME) method. Compared with previous methods, our network can effectively handle the misalignment between the input and the reference without requiring facial priors and learn the aggregated reference set representation in an end-to-end manner. Furthermore, to reconstruct more detailed facial features, we propose a correlation loss that provides a rich representation of the local texture in a controllable spatial range. Experimental results demonstrate that the proposed framework not only has significantly fewer computation cost than recent exemplar-guided methods but also achieves better qualitative and quantitative performance.*

1. Introduction

Numerous psychological and cognitive studies have shown that face perception is one of the most important and specialized aspects of social cognition [17, 35]. The facial regions of a picture tend to draw the attention and interest of observers immediately. Moreover, humans are susceptible to minor changes in familiar faces [38]. Thus, increasing the quality of the face region in images and videos has the potential to significantly enhance the user experience of many social communication applications, *e.g.* real-time video chat, mobile photo booth, *etc*.

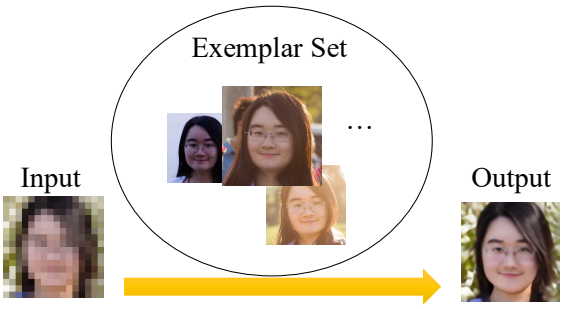

Figure 1: Headshot super-resolution that recovers the lost information in the input using a set of exemplars.

For the above reasons, the machine learning community has widely explored face hallucination [34, 50, 41, 8, 12, 6] as a domain-specific problem of single image super-resolution(SISR) [2, 3, 47], which aims to restore realistic details from a low-resolution (LR) face image to a high-resolution (HR) one. Benefiting from the integration of face structure and identity priors and recent progress in deep neural network designs, it is now possible to generate visually pleasing results even for extremely tiny faces. When the input LR headshot does not contain enough attribute or identity information, using additional references can help to achieve a more faithful reconstruction result. In this paper, we explore a novel method that makes full use of an arbitrarily-sized set of exemplar images to increase the fidelity of headshot image super-resolution.

One core problem is to search the matching regions from references and transfer the corresponding features to the output. Previous methods choose to conduct the global context matching with registration [55], optical flow [45, 60, 30, 15] with a warping [39]. Still, these works assume the exemplars share a similar viewpoint with the LR input [39], which cannot always be guaranteed. Besides, their performance depends on accurate motion estimation and may poorly capture long-range correlations. Other methods [4, 59, 51, 49] conduct an exhaustive patch-wise comparison of LR and reference features, which require a large amount of computation, especially when the reference res-

*This work is done during the author's internship at Meta.
†Affiliated with Meta at the time of this work.

olution is high. In addition, these methods cannot handle inter-patch misalignment or non-rigid deformations. To better use the information of faces from different poses or views, we propose a Reference Feature Alignment module (RFA) that combines optical flow and deformable alignment to find the corresponding information in reference features and align them with the LR content inspired by [31, 7, 9].

In practical applications like smart home cameras or mobile photography, it is possible to acquire many high-resolution images of different views when the user is close to the camera. These images can naturally serve as good exemplars to enhance far-away tiny faces. However, most previous works focus on reference-based super-resolution (RefSR) with one exemplar [59, 51, 39, 30, 15], which is a simplified assumption. To handle a set of exemplars, these methods require an extra step to select the most similar image as the reference according to SIFT [33, 58] or facial landmarks points [29], which is a poor representation of the whole set. [43] devises a framework to process and combine multi-exemplars with a weighted pixel average. Still, it is not robust to the displacement or distortions in reference images, as is our method. To utilize the reference set effectively and efficiently, we propose a Content-conditioned Feature Aggregation module (CoFA) that simplifies the set-to-image RefSR problem to a point-to-point RefSR by aggregating feature maps in a set into a single representation.

Benefiting from the module designs above, our network is end-to-end trainable without requiring other face-specific meta-information. Aiming to generate an SR output with highly-detailed textures, we propose a novel correlation loss inspired by the correlation layer in FlowNet2 [24, 37] to supervise the reconstruction of texture patterns. We compute the pixel-wise correlation across the channel dimension to represent the local textures within a certain window size.

In summary, our contribution is four-fold: (1) we propose a novel headshot super-resolution network that takes advantage of multiple exemplars. Our method is more effective than previous approaches by thoroughly integrating the corresponding information in the exemplar set. It is also computationally efficient since we conduct the matching and transferring in the LR space with careful design; (2) we propose a novel reference feature alignment network to find and align corresponding reference features to the LR content based on flow-guided deformable sampling. We devise a feature aggregation module conditioned on the LR content to explicitly improve the set representation by favoring features that are high in quality and similarity; (3) we propose a novel correlation loss that helps represent the local texture and reconstruct more realistic details; (4) compared with previous approaches, our method achieves state-of-the-art face hallucination performance on the CelebAMask-HQ testset. It also has fewer parameters and computational costs than recent exemplar-guided methods.

2. Related Works

2.1. Reference-based Super-Resolution

Reference-based SR (RefSR) [18] can reconstruct more accurate structures and details benefiting from the reference HR image. The general solution of RefSR includes two steps: searching the matched textures between LR inputs and HR references, and transferring the textures. Some of the previous RefSR approaches choose to align the LR and Ref images with either global registration [55] or optical flow [45, 60]. Other methods choose to match by patches with gradient features [4], or deep features extracted by the CNN [59, 51, 49]. [39] change the feature matching to LR space to reduce computation. [51] introduced the transformer architecture in a cross-scale manner to improve the accuracy of searching and transferring relevant textures. The above works usually include pixel-wise reconstruction loss, perceptual loss [40] and adversarial loss as the objective functions. Zhang *et al*. [58] introduce a Haar wavelet loss and a degradation loss to avoid over-smoothing in final results. Besides, CMSR [13] further expands the reference source from a single image to a pre-built image pool and searches the $k-$nearest patches from the pool. Since these methods exhaustively conduct a patch-wise comparison of LR and reference feature maps, they usually have a high computational cost.

2.2. Face Hallucination

Face hallucination methods can be roughly divided into two categories: blind face hallucination and exemplar-guided restoration. The first category focuses more on integrating face priors in designing the reconstruction network and loss functions: some works include sub-branches for facial landmarks or face structures [61, 41, 5, 54, 25, 52], or face parsing map [12, 11]. Using face structure priors may bring advantages, including the better recovery of the face shape, as reflected by fewer errors on face alignment and parsing. However, the reconstruction results might not look like the same person, especially when the input images contain barely any identifying information. To solve this problem, [56, 23, 20] employ identity information to supervise the training of the reconstruction network. However, these blind reconstruction methods are heavily influenced by the bias within the distribution of training data, and usually fail to generate satisfying results for minority groups.

The second category, exemplar-guided restoration, aims to use another HR image of the same person to improve the visual content quality of the generated images. [30, 15] include a warping sub-network in using the HR guidance, which increases the training steps as well as the computation cost of the network. [29] uses moving least-squares to align the input and guidance images in the feature space and applies AdaIN for feature transfer. It selects a single

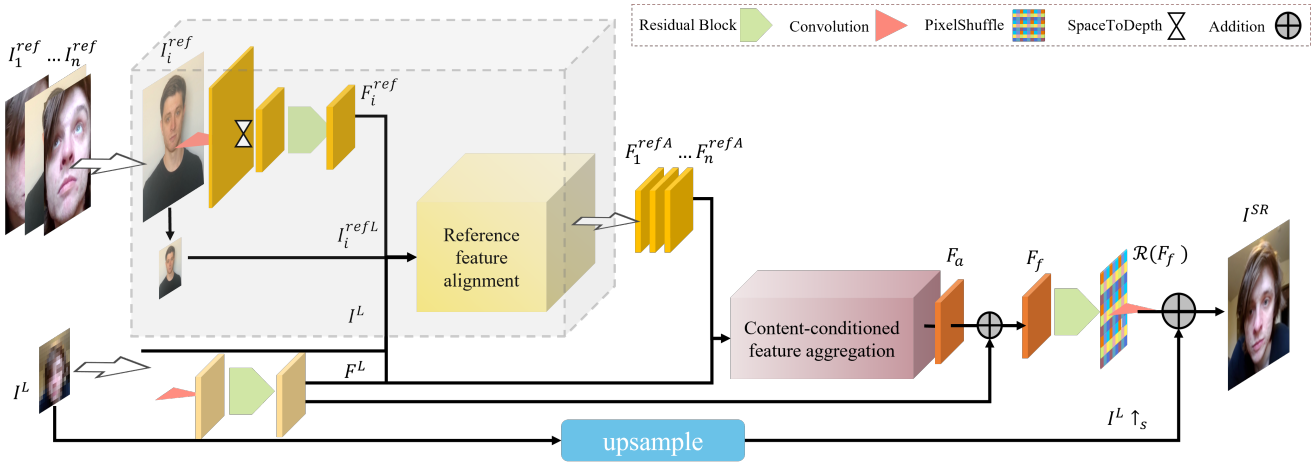

Figure 2: Overview of our **H**eadshot **I**mage Super-Resolution with **M**ultiple **E**xemplars (HIME) framework. Given an input LR image and any number of exemplars, it matches, aligns, and aggregates the features of the reference images conditioned on the input content to reconstruct the SR output.

exemplar from the guidance images, thus cannot fully use the rich information in the guidance face sets. [43] takes a step forward by using multiple exemplars with a weighted pixel average module in the network. But, it cannot handle the large deformation between unaligned faces.

Compared with the approaches above, our method can take full advantage of an unaligned exemplar set as a reference in headshot reconstruction, and our network is end-to-end trainable without requiring face-specific metadata.

3. HIME Framework

Given a low-resolution input I^L and a set of high-resolution headshot images $\mathcal{I}^{ref} = \{I_i^{ref}\}, i = 1, 2, \ldots$ from the same identity, our goal is to generate the corresponding high-resolution image I^{SR}. To efficiently and accurately transfer the matching information from the unaligned reference sets of arbitrary length, we propose the *HIME* framework as illustrated in Figure 2. This framework consists of four main components: *feature extractor*, *reference feature alignment module (RFA)*, *content-conditioned feature aggregation module (CoFA)*, and *HR reconstructor*, as introduced in Sections 3.1, 3.2, 3.3 and 3.4.

We first use an LR feature extractor to get the feature map F^L from I^L and an HR feature extractor to get feature maps $\{F_i^{ref}\}_{i=1}^n$ from the reference set with n HR images. For efficient feature matching and transfer, the reference images and features are converted to the LR space. Then we feed I^L, F^L, $\{I_i^{ref}\}_{i=1}^n$ and $\{F_i^{ref}\}_{i=1}^n$ to the proposed RFA module for alignment. Furthermore, to better utilize the face set information, we use a CoFA module to aggregate the refined features into one. Finally, we reconstruct the HR face image from the aggregated feature map.

3.1. Feature Extractors

We adopt an HR feature extractor and an LR feature extractor to handle images in HR space and LR space, respectively. The HR feature extractor turns the HR reference images into a set of feature maps: $\{F_i^{ref}\}_{i=1}^n$. The RGB images are first converted into a mono-channel feature map since the color information of the reference images is not needed. Then, we adopt a space-to-depth operation to convert the HR feature maps into the same spatial resolution as the input without discarding any information. Next, we apply a convolution layer and k_h residual blocks [22] to extract the HR reference feature maps. The LR feature extractor generates feature maps for the input LR image with a convolutional layer and k_l residual blocks [22].

3.2. Reference Feature Alignment

Given extracted feature maps F^L from the input LR image and $\{F_i^{ref}\}_{i=1}^n$ from the reference images, we want to acquire guiding features that are well-aligned with the contents of the LR image to mitigate any mismatches in view or pose. To achieve this goal, We propose learning a feature alignment function $f(\cdot)$ to directly align the reference feature maps F_i^{ref} as shown in Figure 3. A general form of the alignment function can be formulated as:

$$F_i^{refA} = f(F_i^{ref}, I_i^{refL}, I^L, F^L) = T(F_i^{ref}, \Phi_i), \quad (1)$$

where F_i^{refA} denotes the i-th aligned reference feature, $T(\cdot)$ is the sampling function, and Φ^i is the corresponding sampling parameters. Inspired by the deformable alignment [14, 62] in [44, 42, 48] for spatial and temporal super-resolution, we propose to use deformable sampling functions to implicitly capture the similarities between LR content and reference images. However, the training of de-

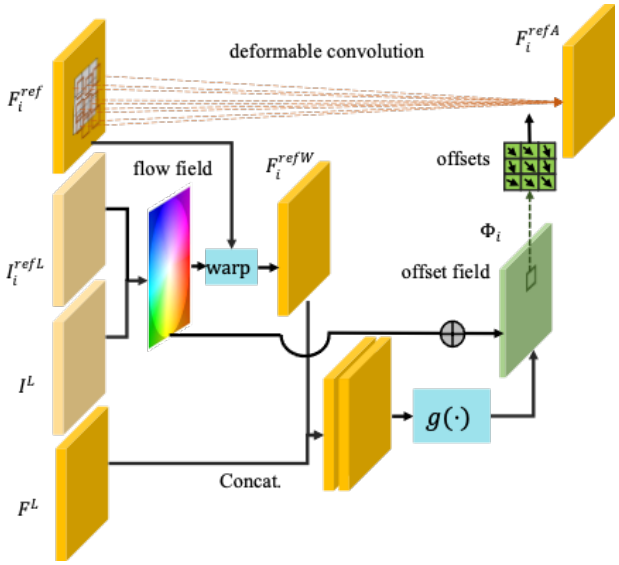

Figure 3: Reference feature alignment (RFA): Optical flow is integrated as part of the offset field to align the reference feature map. Then the aligned features are used to estimate the offset residue. In this way, we can thoroughly exploit the similarities between the LR and reference images.

formable alignment module is hard and full of instability, which might impair the model's final performance. To overcome this issue, we combine the optical flow as guidance.

The offset for the deformable sampling function should be learned based on the correspondences between the reference image and the input LR image, which is very similar to the goal of optical flow. Thus, we directly merge the optical flow into the offset of deformable alignment, and compute the offset residue to further improve the accuracy. We first estimate the optical flow o_i between I^L and I_i^{refL}, and use it to warp the reference features:

$$F_i^{refW} = warp(F_i^{ref}, o_i) \tag{2}$$

Then the warped reference feature is used to predict the offset residual Δp_i, along with the LR feature F^L:

$$\Delta p_i = g([F_i^{refL}, F^L]), \tag{3}$$

where $g(\cdot)$ denotes a general operation of convolution layers for the offset estimation; $[\cdot,\cdot]$ denotes channel-wise concatenation. Then we can acquire the sampling parameters $\Phi_i = o_i + \Delta p_i$. With the flow-guided offset, the sampling function in Equation 1 can be performed with a deformable convolution [14, 62]:

$$F_i^{refA} = T(F_i^{ref}, \Phi_i) = DConv(F_i^{ref}, \Phi_i). \tag{4}$$

We denote the RFA module without optical flow guidance network as the HIME (small), which directly estimating the offset. The network with flow-guided RFA module is demoted as HIME (large).

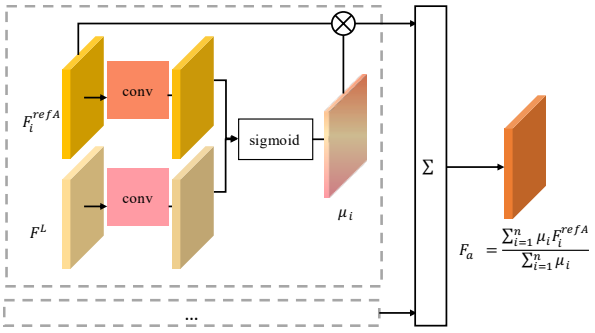

Figure 4: Content-conditioned feature aggregation (CoFA): For each aligned reference feature, we compute a similarity score μ with the input F^L and then aggregate all features F^{refA} with a weighted average.

3.3. Content-conditioned Feature Aggregation

Now we have a set of aligned reference feature maps: $\{F_i^{refA}\}_{i=1}^n$ for the following feature transferring and reconstruction steps. As shown in Figure 4, the CoFA module aims to map this feature map set to a representation with fixed dimension. In this way, the reference image set with a different number of images can be represented in a unified manner. The representation is determined by all items in the set and conditioned on the LR content. Therefore it can be denoted as:$F_a = \mathcal{F}(F_1^{refA}, F_2^{refA}, \ldots, F_n^{refA}|F^L)$, where $\mathcal{F}(\cdot)$ is the aggregation function that maps an arbitrary-sized set to a representation of fixed dimension.

It is challenging to find a proper $\mathcal{F}(\cdot)$ that aggregates features from the whole reference set to obtain an optimized representation. Based on the intuition that references with higher similarity and quality should contribute more to feature transfer, while faces with mismatched features and low-quality features should have less effect on the set representation, we denote $\mathcal{F}(\cdot)$ as:

$$\mathcal{F}(F_1^{refA}, \ldots, F_n^{refA}|F^L) = \frac{\sum_{i=1}^n \mu_i F_i^{refA}}{\sum_i^n \mu_i}, \tag{5}$$

$$\mu_i = \mathcal{S}(F_i^{refA}, F^L), \tag{6}$$

where $\mathcal{S}(\cdot)$ generates a similarity score μ_i for the aligned reference feature map F_i^{refA} that is acquired in the same manner as shown by Equation 4. Therefore, the final representation of the set is a fusion of each feature weighted by its similarity score. For each aligned reference feature F_i^{refA}, the pixel-wise similarity score is calculated as:

$$\mathcal{S}(F_i^{refA}, F^L) = \sigma(g_1(F^{refA})^T g_2(F^L)), \tag{7}$$

where $\sigma(\cdot)$ is sigmoid function that is used for bounding the outputs to the range $[0, 1]$ and stabilizing the gradient propagation; and $g_1(\cdot)$ and $g_2(\cdot)$ denotes general convolution layers. The similarity score can also be regarded as an attention mask conditioned on the input content.

Finally, the summation F_a and LR feature map is sent to HR image reconstruction: $F_f = F_a + F^L$. The similarity computation and weighted aggregation steps are parameter-free. Thus, the CoFA module is light-weighted by design.

3.4. High-Resolution Image Reconstruction

The HR reconstruction module takes the fused feature F_f as input and generates the residual of our target HR output. It is composed of k_r stacked residual blocks [22] for learning deep features and a sub-pixel upsampling module with PixelShuffle [21] initialized using the ICNR method as in [1, 47]. To encourage the network to focus on learning high-frequency information that is not present in the LR input, we introduce a long-range skip connection to form the final SR output: $I^{SR} = I^L \uparrow_s + \mathcal{R}(F_f)$, where $\uparrow$ denotes the bicubic upscaling operation and s denotes the scale factor; $\mathcal{R}(\cdot)$ denotes the reconstruction operations as described above. Allowing the low-frequency information in the LR input to bypass the reconstruction network lowers the difficulty of reconstruction learning and accelerates the convergence of the optimization process.

Since the input and reference images are highly related in the face domain, our model can simultaneously learn the feature alignment and similarity score with only supervision from the HR ground truths through the end-to-end training.

4. Correlation Loss

Motivation. The commonly used pixel-wise reconstruction losses inevitably lead to over-smoothing of outputs and don't match the human visual perception of natural images [26], since they fail to capture the underlying local relationships between pixels. While the perceptual loss [40] and style loss [19] have been introduced to provide more perception-oriented supervision, they require a pretrained network from another high-level vision task, and are not versatile for representing textures of very high-resolution images due to the limits of training data. To effectively represent the local texture patterns of different scales in a controllable manner, we devise the correlation loss. It first builds a correlation map from the correlation between the center pixel and its neighbors to represent the spatial patterns. Thus, matching the correlation map can help the network reconstruct more realistic details and improve the perceptual quality of the output images.

Design of Correlation Loss. As shown in Figure 5, each image I can be represented by a 3D tensor of size (C, H, W), where C is the number of channels and (H, W) denotes the spatial resolution. We first subtract the mean of each channel to center the data around 0. For a given pixel $I(x, y)$, we calculate its inner product with the neighboring pixels $I(x-i, y-j)$ as well as itself within a $k \times k$ window:

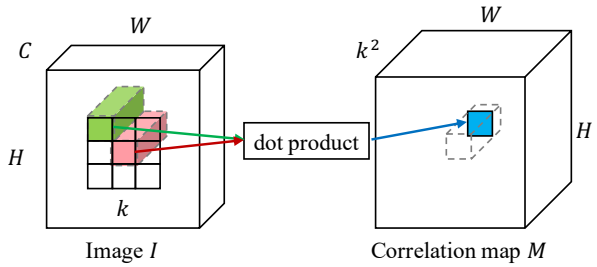

Figure 5: Illustration of the proposed correlation loss. The correlation operator is used for both generated and ground-truth images. Then we take the corresponding output correlation maps to calculate the correlation loss.

$$cor(i, j, x, y) = \frac{1}{k^2}\langle I(x, y), I(x-i, y-j)\rangle, \quad (8)$$

where $\langle , \rangle$ denotes inner product, $i, j \in \lfloor -\frac{k+1}{2}, \frac{k+1}{2} \rfloor$, and $\frac{1}{k^2}$ is for normalization. k is the maximal displacement for computing the local correlation. As a result, we can acquire a correlation map M_{cor} of size $(k \times k, H, W)$. The correlation loss is the distance between the correlation maps from the ground truth HR and the generated SR images:

$$L_{cor} = ||M_{cor}^{HR} - M_{cor}^{SR}||. \quad (9)$$

In our implementation, we adopt the $L1-$distance for this loss term. A larger window size k can encode more information while quadratically increasing the computational cost. Thus, we define the dilated correlation following the same manner as the dilated convolution [53]. By increasing the dilation factor d, we can enlarge the correlation window from $k \times k$ to $(kd - d + 1) \times (kd - d + 1)$.

Visualizing Correlation Maps. To better understand the correlation operation, we visualize the correlation maps of the HR image with different correlation kernel window sizes $k \in \{3, 5, 7\}$. In Figure 6, we observe that the correlation map encodes the original image based on the local textures. In each correlation map, the blue areas correspond to the regions with more high-frequency features, like furs and the background, regardless of the color difference. While the red regions are more smooth, *e.g.*, the brightest and darkest part of the fur. With the increase of window size k, the correlation operator perceives and encodes features within a broader area, and thus looks more coarse-grained in the visualized results.

5. Experiment

5.1. Implementation Details

In our implementation, $k_l = 5$, $k_h = 3$, and $k_r = 20$ residual blocks are used in LR feature extraction, HR feature extraction, and HR image reconstruction modules, respectively. For each LR input, we randomly select three

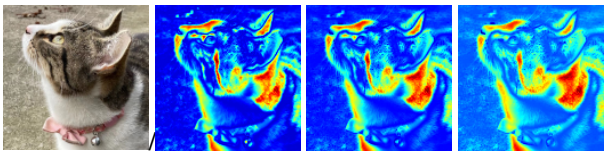

Image $k = 3$ $k = 5$ $k = 7$

Figure 6: Visualization of correlation maps of different window sizes. The image is 128×128-resolution.

different HR images to build the reference set during training. We adopt SPyNet [36] as the optical flow estimator in HIME (large). More details can be found in the supplementary materials.

Objective Function. For a fair comparison with previous methods, we train two types of models: (1) reconstruction-oriented models HIME$_{rec}$ with the pixel-wise reconstruction loss L_{rec} and our proposed correlation loss L_{cor}. W the Charbonnier penalty function [26] as the loss term for pixel-wise reconstruction to optimize our framework: $L_{rec} = \sqrt{||I^{HR} - I^{SR}||^2 + \epsilon^2}$, where I^{HR} denotes the ground-truth HR frame, and ϵ is empirically set to 1×10^{-3}. (2) perception-oriented models HIME$_P$ include L_{rec}, L_{cor}, and the adversarial loss L_{adv}, the perceptual loss L_{per}:

$$\mathcal{L}_P = \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv} + \lambda_{per} L_{per} + \lambda_{cor} L_{cor}, \quad (10)$$

where λs are the weights for each loss term.

Datasets. CelebAMask-HQ is used as the training and evaluation datasets [27], including over 30,000 high-resolution headshots selected from the CelebA dataset [32]. We acquire the identity information from the original CelebA dataset and remove 3,300 out of 6,217 identities with < 4 images, which are not enough to construct a set of multiple references. The remaining identities are randomly split into a training set and an evaluation set, including 2,600 and 287 identities, respectively. We generate images of different scales by bicubic downsampling with factor $= s$.

Evaluation Metrics. We adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [46] metrics to evaluate the reconstruction performance on all RGB channels. We also compare the perceptual quality with LPIPS [57]. To measure the efficiency of the different methods, we report the model parameters and computational cost for each setting.

5.2. Comparison to the State of the Art

We evaluate the performance of our HIME network under the $4\times$ and $8\times$ upsampling setting following the previous approaches. For $4\times$ upscale, we compare two SOTA RefSR methods: SRNTT [59][1] and TTSR [51], and three recent face restoration method SPARNet [10], PSFR-GAN [11] and DFDNet [28]. We did not test DFDNet [28] on the 32×4 setting since its face and landmark detectors cannot handle such tiny faces. For $8\times$ upsampling, we compare our method with five face hallucination methods: PFSR [6], FSRNet [12][2], GWAInet [15], SPARNet [10] and PSFR-GAN [11]. Quantitative results are shown in Table 1.

(LR, s)	Methods	PSNR	SSIM	LPIPS	Params (M)	GMACs
$(32, 4)$	Bicubic	25.64	0.7752	0.3229	-	-
	SRNTT [59]	28.02	0.8434	0.0682	6.30	36.47
	TTSR [51]	27.31	0.8346	0.0633	6.73	26.62
	SPARNet [10]	20.50	0.6118	0.1617	85.73	45.25
	PSFR-GAN [11]	25.47	0.7709	0.0981	67.05	117.84
	HIME$_{rec}$ (small)	29.11	0.8794	0.1136	**0.87**	**1.86**
	HIME$_P$ (small)	27.16	0.8269	0.0464	**0.87**	**1.86**
	HIME$_{rec}$ (large)	**29.23**	**0.8817**	0.1102	9.23	6.06
	HIME$_P$ (large)	27.05	0.8224	**0.0461**	9.23	6.06
$(64, 4)$	Bicubic	28.40	0.8169	0.2860	-	-
	SRNTT [59]	30.41	0.8552	0.0906	6.30	145.89
	TTSR [51]	29.87	0.8484	0.0851	6.73	106.48
	SPARNet [10]	23.26	0.6990	0.1341	85.73	180.99
	PSFR-GAN [11]	26.62	0.7685	0.1039	67.05	161.89
	DFDNet [28]	21.55	0.6587	0.1581	133.34	601.04
	HIME$_{rec}$ (small)	31.24	0.8785	0.1611	**0.87**	**7.48**
	HIME$_P$ (small)	29.06	0.8262	**0.0633**	**0.87**	**7.48**
	HIME$_{rec}$ (large)	**31.28**	**0.8789**	0.1600	9.23	24.24
	HIME$_P$ (large)	29.16	0.8272	0.0641	9.23	24.24
$(16, 8)$	Bicubic	21.83	0.5929	0.5247	-	-
	PFSR[6]	21.44	0.5778	0.2065	10.08	8.97
	FSRNet [12]	20.03	0.5749	0.2865	15.52	3.20
	GWAINet [15]	21.96	0.5844	0.2056	4.29	6.55
	SPARNet [10]	19.00	0.5022	0.2576	85.73	45.25
	PSFR-GAN [11]	22.05	0.6102	0.2062	67.05	117.84
	HIME$_{rec}$ (small)	24.54	0.7411	0.2433	**0.90**	**0.49**
	HIME$_P$ (small)	22.45	0.6338	**0.1297**	**0.90**	**0.49**
	HIME$_{rec}$ (large)	**24.68**	**0.7467**	0.2361	9.26	4.49
	HIME$_P$ (large)	23.35	0.6744	0.1313	9.26	4.49

Table 1: Quantitative comparison of our results and other SOTA methods. The best results are shown in **bold**.

From Table 1, we can learn the following facts: (1) reference-based SR methods, like SRNTT, TTSR and our HIME, demonstrate better performance than other non-reference approaches on both distortion-oriented metrics and perception-oriented metrics, which validate that using references can improve the SR fidelity. Our network outperforms the other result by 1.21/1.09 dB on $(32, 4)$, and 0.87/0.83 dB on $(64, 4)$; (2) Although SRNTT and TTSR have fewer parameters than other compared methods, their computational costs are relatively high due to the exhaustive search during feature matching. With the learnable feature extractors, our small model is over $7\times$ smaller than SRNTT and TTSR. The reference feature alignment in LR space make our network have 14.3 and 4.39 $\times$ fewer GMACs than TTSR. For the $(16, 8)$ setting, we can observe that our method performs well even under the very challenging $8\times$ upsampling setting.

The visual results on the DFDC dataset [16] are shown in Figure 7, which validates our observations above. RefSR methods like SRNTT, TTSR and ours can generate more robust and visually pleasing results. For the GAN-based face enhancement methods SPARNet and PFSR-GAN, while

[1] PyTorch implementation: https://github.com/S-aiueo32/srntt-pytorch

[2] PyTorch implementation: https://github.com/cydiachen/FSRNET_pytorch

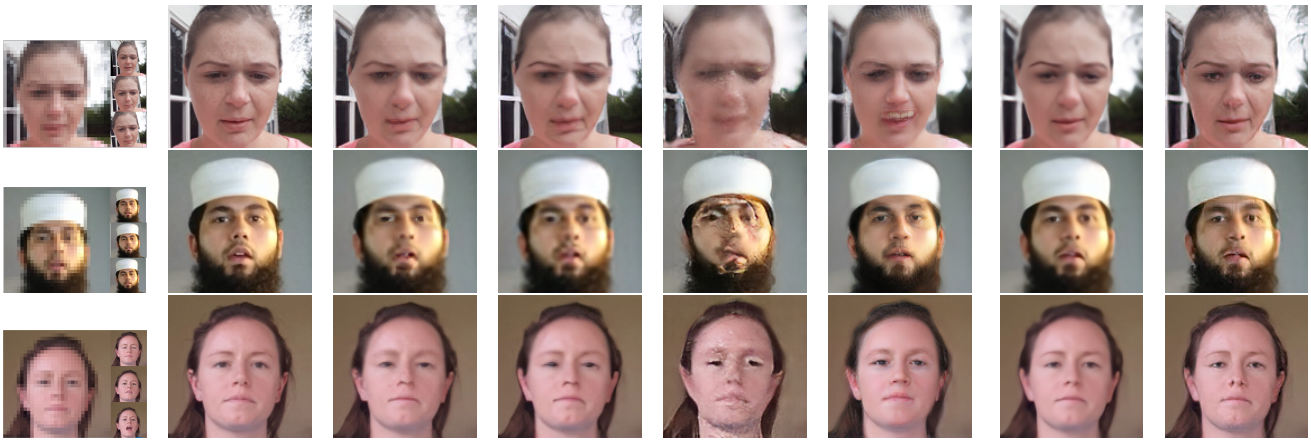

Figure 7: Qualitative comparison with SOTA methods for $4\times$ upscale setting. Input resolution: 32×32.

their results are rich in details, sometimes they fail on tiny faces with deformations.

5.3. Ablation Study

We perform comprehensive ablation studies to further demonstrate the effectiveness of each modules in our network, the influence of exemplars and the correlation loss. Experiments below are conducted under $8\times$ upscale with input size 16×16 images, if not specified otherwise.

Effectiveness of Reference Feature Alignment. To investigate the proposed RFA module, we compare three models: (a), (b), and (c), where (a) replaces the deformable convolution in the RFA module with common convolution that does not have the capability of feature alignment, and (b) is our small model by removing the optical flow guidance, and directly estimate the offset with F_i^{ref} and F^L, (c) is our large model as illustrated in Section 3.2

Methods	PSNR↑	SSIM↑	LPIPS↓
Conv	24.33	0.7311	0.2605
Dconv	24.54	0.7411	0.2433
Dconv-flow	24.68	0.7467	0.2361

Table 2: Ablation study of feature alignment methods.

From Table 2, we can see that adopting the deformable alignment brings up the performance on all metrics compared with using the common convolution. And the flow-guided deformable alignment can further improve the performance. The results demonstrate that our RFA module can better match the features between the LR content and the references and is more robust to the misalignment and distortion. Our network conducts the offset computation and feature matching in the LR space, achieving a better performance while reducing the computational cost.

Set Feature Aggregation. To validate the effect of our proposed feature aggregation mechanism in the CoFA module, we compare three different models: (a) averages the features without content conditioning, (b) aggregates the features by max-pooling across the set, and (c) is our proposed aggregation method weighted by the learned content similarity. The quantitative results are shown in the Tabel 3.

Methods	PSNR↑	SSIM↑	LPIPS↓
Average	22.120	0.6350	0.4332
Max-pool	22.118	0.6349	0.4331
CoFA	24.381	0.7339	0.2533

Table 3: Ablation study of feature aggregation methods.

From Table 3, we can see that the model with our content-conditioned feature aggregation module outperforms the average and max pooling by over 2 dB in terms of PSNR. Adopting the CoFA module greatly improves performance on all metrics, which indicates that our designed module can extract a better set representation, helping to restore the LR information and enhance the output quality.

Effect of Multiple Exemplars. To validate whether using an exemplar set can improve the face super-resolution result, we conduct the following experiments: (a) non-ref: a baseline SR network without references and removing the HR matching and aggregation modules, (b) training and testing with one reference image and (c) with three reference images. From the results in Table 4, we can observe that using references significantly increases the PSNR by 0.49 dB while using multiple references further improves it by 0.19 dB. Such improvements also apply to the SSIM and LPIPS. These results verify that our model can benefit from the rich information in the exemplar set, and can effectively utilize the corresponding features to improve the output quality.

Influence of Exemplar Similarity. Our method has the potential to be applied on video calling, where the close-to-camera headshots can be used to enhance the far-away ones when zooming in. For this scenario, we recorded several video calls from ourselves and collected over 5,000 frames

Num of Ref	PSNR↑	SSIM↑	LPIPS↓
0	23.84	0.7088	0.3440
1	24.35	0.7318	0.2572
3	24.54	0.7409	0.2433

Table 4: Ablation study of multiple exemplars by changing the number of references during training and testing.

to verify the influence of the temporal gap. Intuitively, with the increase of interval j, the Ref is less similar to the LR inputs due to the motion in natural videos. We downsample these frames $4\times$ to construct LR inputs, and pick an HR image as Ref every j frame. We also experiment on using a blank image as a reference, which does not provide any similar features. From Table 5, we can observe that the performance decreases with the larger temporal interval and less similarity, and gracefully descends to a lower bound. Still, using Refs shows better results than blank Ref in terms of PSNR and SSIM.

Interval j	30	60	120	Blank Ref
PSNR	37.34	37.24	37.12	36.81
SSIM	0.9250	0.9241	0.9227	0.9207

Table 5: Influence of temporal gap (interval) between input and reference images.

Effect of Correlation Loss. To justify the effectiveness of correlation loss, we experimentally compare different configurations of HIME in Table 6. We consider the following models: (a) reconstruction loss only; (b) reconstruction loss + correlation loss; (c) multiple losses in GAN training (without correlation loss); (d) correlation loss + (c).

From Table 6, by comparing the first two rows, we can observe that introducing the correlation loss slightly decreases the PSNR. However, it improves the structural and perceptual metrics SSIM and LPIPS, which demonstrates that the proposed correlation loss benefits the reconstruction of local textures. Comparing the last two rows, training with the correlation loss greatly leverages the perception-oriented model's performance on all metrics, which further validates the effectiveness of the correlation loss as perception-oriented supervision.

Figure 8 shows the performance of HIME for different correlation window sizes $k \in \{1, 3, 5, 7, 9\}$, where $k = 1$ degrades to the common $L1$ loss of the squared pixel values. We conduct two types of experiments: (a) training

Methods	PSNR↑	SSIM↑	LPIPS↓
L_{rec}	24.38	0.7339	0.2533
$L_{rec} + L_{cor}$	24.35	0.7346	0.2437
L_P w/o L_{cor}	22.44	0.6204	0.1543
L_P w/ L_{cor}	23.28	0.6673	0.1389

Table 6: Effectiveness of our proposed correlation loss.

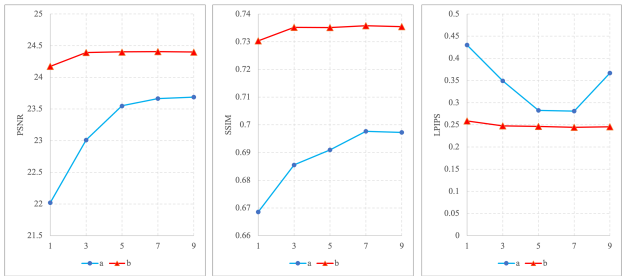

Figure 8: Effect of correlation window size k on output quality in terms of PSNR, SSIM, and LPIPS: (a) training with L_{cor} only, (b) fine-tuning with both L_{rec} and L_{cor}.

with correlation-loss only (plotted in blue), (b) fine-tuning with both L_{rec} and L_{cor} (plotted in red). Viewing the blue plots, we can observe that with the growth of k, the model performs better in terms of PSNR and SSIM. These results demonstrate that the correlation map itself is a good representation of the RGB image. With a larger window size, the correlation map can encode more information. Still, such improvement becomes more marginal when k is large enough. When $k = 9$, the LPIPS even increases. As for the red plots, we can see a similar trend: when $k \geq 3$, the improvement on PSNR and SSIM is very trivial. These results indicate that for a certain scale, there exists a range of k that work best in representing the local patterns. Within this range, the LPIPS scores keep decreasing with the increase of k. It implies that the correlation loss is more like perception-oriented supervision, which validates our description in Section 4.

6. Conclusion and Future Work

In this paper, we propose an effective framework for headshot image super-resolution with multiple exemplars without face structure priors. To achieve this, we introduce a reference feature alignment module to search and align corresponding features to the LR content. To construct an optimized set representation, we propose a feature aggregation network conditioned on the input content. With such a design, our network can learn to fully utilize the rich information in the exemplar set and be robust to misalignment and deformations. Furthermore, we propose a correlation loss that supervises the reconstruction of local textures with correlation maps. We believe that our new **H**eadshot **I**mage Super-Resolution with **M**ultiple **E**xemplars network (HIME) provides a novel idea to efficiently utilize a set of data for the reference-based super-resolution and face hallucination task. In future works, we will explore other aggregation methods to generate a better set representation with the aid of face priors. In addition, we will further validate the effectiveness of the correlation loss as generic supervision for other low-level tasks, *e.g.* image denoising, video frame interpolation, style transfer, *etc*.

References

[1] Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

[2] Jan Allebach and Ping Wah Wong. Edge-directed interpolation. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pages 707–710. IEEE, 1996.

[3] Clayton Brian Atkins, Charles A Bouman, and Jan P Allebach. Optimal image scaling using pixel classification. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 3, pages 864–867. IEEE, 2001.

[4] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014.

[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*, pages 109–117, 2018.

[6] 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 the European Conference on Computer Vision*, pages 185–200, 2018.

[7] Zhuojun Cai, Xiang Tian, Ze Chen, and Yaowu Chen. Space-time super-resolution with motion-perceptive deformable alignment. *Journal of Electronic Imaging*, 30(3):033020, 2021.

[8] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 690–698, 2017.

[9] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022.

[10] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K Wong. Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231, 2020.

[11] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. *arXiv preprint arXiv:2009.08709*, 2020.

[12] 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*, pages 2492–2501, 2018.

[13] Shuguang Cui. Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation. 2020.

[14] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[15] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[16] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[17] Martha J Farah, Kevin D Wilson, Maxwell Drain, and James N Tanaka. What is "special" about face perception? *Psychological Review*, 105(3):482, 1998.

[18] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002.

[19] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[20] Klemen Grm, Walter J Scheirer, and Vitomir Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29:2150–2165, 2019.

[21] Manuel Guizar-Sicairos, Samuel T Thurman, and James R Fienup. Efficient subpixel image registration algorithms. *Optics Letters*, 33(2):156–158, 2008.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Chih-Chung Hsu, Chia-Wen Lin, Weng-Tai Su, and Gene Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28:6225–6236, 2019.

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017.

[25] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019.

[26] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017.

[27] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020.

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

[31] Jiayi Lin, Yan Huang, and Liang Wang. Fdan: Flow-guided deformable alignment network for video super-resolution. *arXiv preprint arXiv:2105.05640*, 2021.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015.

[33] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, 1999.

[34] Jeong-Seon Park and Seong-Whan Lee. An example-based face hallucination method for single-frame, low-resolution facial images. *IEEE Transactions on Image Processing*, 17:1806–1816, 2008.

[35] Kimberly A Quinn and C Neil Macrae. The face and person perception: Insights from social cognition. *British Journal of Psychology*, 102(4):849–867, 2011.

[36] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proc. CVPR*, pages 4161–4170, 2017.

[37] Fitsum Reda, Robert Pottorff, Jon Barker, and Bryan Catanzaro. flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks. `https://github.com/NVIDIA/flownet2-pytorch`, 2017.

[38] Gillian Rhodes and Jim Haxby. *Oxford Handbook of Face Perception*. Oxford University Press, 2011.

[39] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2020.

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

[41] Yibing Song, Jiawei Zhang, Shengfeng He, Linchao Bao, and Qingxiong Yang. Learning to hallucinate face images via component generation and enhancement. *arXiv preprint arXiv:1708.00223*, 2017.

[42] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[43] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Multiple exemplars-based hallucination for face super-resolution and editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[44] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[45] Yuwang Wang, Yebin Liu, Wolfgang Heidrich, and Qionghai Dai. The light field attachment: Turning a DSLR into a light field camera using a low budget camera ring. *IEEE Transactions on Visualization and Computer Graphics*, 23:2357–2364, 2016.

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[47] Xiaoyu Xiang, Qian Lin, and Jan P Allebach. Boosting high-level vision with joint compression artifacts reduction and super-resolution. *arXiv preprint arXiv:2010.08919*, 2020.

[48] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3370–3379, 2020.

[49] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *European Conference on Computer Vision*, pages 230–245, 2020.

[50] Chih-Yuan Yang, Sifei Liu, and Ming-Hsuan Yang. Structured face hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1099–1106, 2013.

[51] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[52] Yu Yin, Joseph Robinson, Yulun Zhang, and Yun Fu. Joint super-resolution and alignment of tiny faces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12693–12700, 2020.

[53] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

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

[55] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, 22:4865–4878, 2013.

[56] 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)*, pages 183–198, 2018.

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

[58] Yulun Zhang, Zhifei Zhang, Stephen DiVerdi, Zhaowen Wang, Jose Echevarria, and Yun Fu. Texture hallucination for large-factor painting super-resolution. *arXiv preprint arXiv:1912.00515*, 2019.

[59] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[60] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision*, pages 88–104, 2018.

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

[62] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.