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 an efficient Headshot Image Super-Resolution with Multiple Exemplars 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, 37]. 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 [40]. 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.

For the above reasons, the machine learning community has widely explored face hallucination [6, 8, 12, 36, 43, 52] as a domain-specific problem of single image super-resolution (SISR) [2, 3, 49], 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 [57], optical flow [15, 31, 47, 62] with a warping [41]. Still, these works assume the exemplars share a similar viewpoint with the LR input [41], which cannot always be guaranteed. Besides, their performance depends on accurate motion estimation and may poorly capture long-range correlations. Other methods [4, 51, 53, 61] conduct an exhaustive patch-wise comparison of LR and reference features, which require a large amount of computation, especially when the reference resolution 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 [7, 9, 32].

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 [15, 31, 41, 53, 61], 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 [35, 60] or facial landmarks points [30], which is a poor representation of the whole set. [45] 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, 39] 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 [57] or optical flow [47, 62]. Other methods choose to match by patches with gradient features [4], or deep features extracted by the CNN [51, 53, 61]. [41] change the feature matching to LR space to reduce computation. [53] 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 [42] and adversarial loss as the objective functions. Zhang et al. [60] 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 [5, 26, 43, 54, 56, 63], or face parsing map [11, 12]. 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, [20, 23, 58] 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. [15, 31] include a warping sub-network in using the HR guidance, which increases the training steps as well as the computation cost of the network. [30] 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 exemplar from the guidance images, thus cannot fully use the rich information in the guidance face sets. [45] 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 \( l^L \) and a set of high-resolution headshot images \( I^{ref} = \{I_{i}^{ref}\}, i = 1, 2, \ldots \) 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 \( \{I_{i}^{ref}\}_{i=1}^n \) 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 \( F_L, F_L, \{I_{i}^{ref}\}_{j=1}^n \) and \( \{I_{i}^{ref}\}_{j=1}^n \) 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 monochannel 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 $k_r$ 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 $k_r$ residual blocks [22].

### 3.2 Reference Feature Alignment

Given extracted feature maps $F^L$ from the input LR image and $\{F^{ref}_i\}_{i=1}^n$ 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 $F^{ref}_i$ as shown in Figure 3. A general form of the alignment function can be formulated as:

$$F^{refA}_i = f(F^{ref}_i, F^{refL}_i, l^L, F^L) = T(F^{ref}_i, \Phi_i),$$  

(1)

where $F^{refA}_i$ denotes the $i$-th aligned reference feature, $T(\cdot)$ is the sampling function, and $\Phi_i$ is the corresponding sampling parameters. Inspired by the deformable alignment [14, 64] in [44, 46, 50] 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 deformable 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 $l^L$ and $l^{refL}_i$, and use it to warp the reference features:

$$F^{refW}_i = warp(F^{ref}_i, o_i)$$  

(2)

Then the warped reference feature is used to predict the offset residual $\Delta p_i$, along with the LR feature $F^L$:

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

(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, 64]:

$$F^{refA}_i = T(F^{ref}_i, \Phi_i) = DConvo(F^{ref}_i, \Phi_i).$$  

(4)

Besides, we construct the HIME (small) by removing the optical flow network in RFA and directly estimating the offset without flow guidance, as shown in Fig. 3 (a).

### 3.3 Content-conditioned Feature Aggregation

Now we have a set of aligned reference feature maps: $\{F^{refA}_i\}_{i=1}^n$ 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: $F_a = \mathcal{F}(F^{refA}_1, F^{refA}_2, \ldots, F^{refA}_n | \mu_L)$, 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:

$$\mathcal{F}(F^{refA}_1, \ldots, F^{refA}_n | \mu_L) = \frac{\sum_{i=1}^{n} \mu_i F^{refA}_i}{\sum_{i} \mu_i}.$$  

(5)
where \( S(\cdot) \) generates a similarity score \( \mu_i \) 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:

\[
S(F_i^{refA}, F) = \sigma(g_1(F_i^{refA})^T g_2(F)),
\]

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 \( k_f \) 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, 49]. 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: \( I^{SR} = I^L \uparrow_s + R(F_f) \), where \( \uparrow_s \) denotes the bicubic upsampling operation and \( s \) denotes the scale factor; \( R(\cdot) \) 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. We use the Charbonnier penalty function [27] 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 \( \epsilon \) is empirically set to \( 1 \times 10^{-3} \).

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 [27], since they fail to capture the underlying local relationships between pixels. While the perceptual loss [42] 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 substract 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:

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

where \(\langle , \rangle\) denotes inner product, \(i, j \in [-\frac{k + 1}{2}, \frac{k + 1}{2}]\), 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:

\[
L_{cor} = ||M^{HR}_{cor} - M^{SR}_{cor}||.\tag{9}
\]

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 [55]. 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.

![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.](Image 324x649 to 379x704)

![Figure 6: Visualization of correlation maps of different window sizes. The image is 128 × 128-resolution.](Image 439x648 to 494x703)
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 $\times 4$. For each LR input, we randomly select three different HR images to build the reference set during training. In the evaluation stage, we randomly select reference images to form an evaluation list. This list is applied to all evaluated methods for a fair comparison.

**Evaluation Metrics.** We adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [48] metrics to evaluate the reconstruction performance on all RGB channels. We also compare the perceptual quality with LPIPS [59]. 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 ReSR methods: SRNTT [61] and TTSR [53], and three recent face restoration method SPARNet [10], PSFR-GAN [11] and DFDNet [29]. We did not test DFDNet [29] on the $32 \times 4$ 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: FSRNet [6], FSRNet [12], GWANet [15], SPARNet [10] and PSFR-GAN [11]. Quantitative results are shown in Table 1.

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

| (LR, x) | Methods | PSNR | SSIM | LPIPS | Params (M) | GMAC's |
|---------|---------|------|------|-------|------------|--------|
| (32, 4) | Bicubic | 25.64 | 0.7752 | 0.3229 | - | - |
|         | SRNTT [61] | 28.02 | 0.8434 | 0.0682 | 6.30 | 36.47 |
|         | TTSR [53] | 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 | 0.7769 | 0.0981 | 67.05 | 117.84 |
|         | HIME (small) | 29.11 | 0.8794 | 0.1136 | 0.87 | 1.86 |
|         | HIME (small) | 27.16 | 0.8269 | 0.0464 | 0.87 | 1.86 |
|         | HIME (large) | 29.23 | 0.8817 | 0.1002 | 9.23 | 6.06 |
|         | HIME (large) | 27.05 | 0.8224 | 0.0461 | 9.23 | 6.06 |
| (64, 4) | Bicubic | 28.40 | 0.8169 | 0.2860 | - | - |
|         | SRNTT [61] | 30.41 | 0.8552 | 0.0906 | 6.30 | 145.89 |
|         | TTSR [53] | 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 | 117.84 |
|         | HIME (small) | 31.24 | 0.8785 | 0.1611 | 0.87 | 7.48 |
|         | HIME (small) | 29.06 | 0.8262 | 0.0633 | 0.87 | 7.48 |
|         | HIME (large) | 31.28 | 0.8789 | 0.1600 | 9.23 | 24.24 |
|         | HIME (large) | 29.16 | 0.8272 | 0.0641 | 9.23 | 24.24 |
| (16, 8) | Bicubic | 21.83 | 0.5929 | 0.5247 | - | - |
|         | PSFR-GAN [6] | 21.44 | 0.5778 | 0.2065 | 10.08 | 8.97 |
|         | FSRNet [12] | 20.03 | 0.5749 | 0.2865 | 15.52 | 3.20 |
|         | GWANet [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] | 22.95 | 0.6102 | 0.2062 | 67.05 | 117.84 |
|         | HIME (small) | 24.34 | 0.7411 | 0.2433 | 0.90 | 0.49 |
|         | HIME (small) | 22.45 | 0.6338 | 0.1297 | 0.90 | 0.49 |
|         | HIME (large) | 24.68 | 0.7467 | 0.2361 | 9.26 | 4.49 |
|         | HIME (large) | 23.55 | 0.6744 | 0.1513 | 9.26 | 4.49 |

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 makes 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. ReSR methods like SRNTT, TTSR, and ours can generate more robust and visually pleasing results. For the GAN-based face enhancement methods SPARNet and PSFR-GAN, while their results are rich in details, sometimes they fail on tiny faces with deformations.

### 5.3 Ablation Study

We perform a comprehensive ablation study to further demonstrate the effectiveness of different modules in our network and the correlation loss. We also discuss the influence of face chirality under different data augmentation strategies. All experiments below are conducted under the same setting: $8\times$ upscale with input size $16 \times 16$ images.

**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 $F^\text{ref}_i$ and $F^i_\text{opt}$, (c) is our large model as illustrated in Section 3.2.

**Table 2: Ablation study of feature alignment methods.**

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

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 Table 3.

Table 3: Ablation study of feature aggregation methods.

| 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 |

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 Reference Images 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 to verify the influence of the temporal gap. We downsample these frames $4\times$ to construct LR inputs, and pick an HR image as Ref every $j$ frame. Intuitively, with the increase of interval $j$, the Ref is less similar to the LR inputs due to the motion in natural videos. 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.

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

| Interval | PSNR | SSIM |
|----------|------|------|
| 30       | 37.34 | 0.9250 |
| 60       | 37.24 | 0.9241 |
| 120      | 37.12 | 0.9227 |
| Blank    | 36.81 | 0.9207 |

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
Table 6: Effectiveness of our proposed correlation loss.

| 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 |

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 $L_{cor}$.

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 $L_1$ loss of the squared pixel values. We conduct two types of experiments: (a) training 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 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 $k \geq 9$, the LPIPS even decreases. As for the red plots, we can see a similar trend: when $k \geq 3$, the improvement in 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.

Face Chirality. Typical human faces contain a variety of asymmetries, [33] brings up the visual chirality in faces and the distribution bias in public face datasets. Here we compare the influence of such asymmetries in headshot RefSR to answer the following questions: 1) Does the mismatch between input LR and references matter? 2) Does the bias in the training set influence the reconstruction performance?

Table 7 shows our experimental results of changing the augmentation: (a) no augmentation; (b) randomly horizontal-flip the LR or Ref images, but not both for a given pair, which introduces the face view mismatch; (c) randomly horizontal-flip both the LR and reference images, which balances the number of left and right faces without introducing mismatches. By comparing (a) and (b), we can observe that the PNSR drops by 0.08 and the SSIM drops by 0.006, respectively, which indicates that training with mismatched views of faces would impair the model’s performance. Compared with (a), (c) performs slightly better in terms of the PSNR and SSIM, which demonstrates that the proper augmentation improves the performance by mitigating the distribution bias in the dataset.

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 Headshot Image Super-Resolution with Multiple Exemplars 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. 2017. 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. 1996. Edge-directed interpolation. In Proceedings of 3rd IEEE International Conference on Image Processing, Vol. 3. IEEE, 707–710.
[3] Clayton Brian Atkins, Charles A Bouman, and Jan P Allebach. 2001. Optimal image scaling using pixel classification. In Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), Vol. 3. IEEE, 864–867.
[4] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. 2014. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In 2014 IEEE International Conference on Computational Photography (ICCP). IEEE, 1–10.
[5] Adrian Bulat and Georgios Tzimiropoulos. 2018. 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. 169–177.
[6] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. 2018. 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. 185–200.
[7] Zhiyuan Cai, Xiang Tian, Ze Chen, and Yaqiu Chen. 2021. Space-time super-resolution with motion-perceptive deformable alignment. Journal of Electronic Imaging 30, 3 (2021), 033020.
[8] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. 2017. Attention-aware face hallucination via deep reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 690–698.
[9] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2021. BasicVSR++. Improving Video Super-Resolution with Enhanced Propagation and
[61] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. 2019. Image super-resolution by neural texture transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7982–7991.

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

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

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

Lucas Young1, Federico Perazzi1, Rakesh Ranjan1, Amit Kumar1, Andrea Colaco1, Jan Allebach2.