Arbitrary-Resolution and Arbitrary-Scale Face Super-Resolution with Implicit Representation Networks

Yi Ting Tsai, Yu Wei Chen, Hong-Han Shuai, and Ching-Chun Huang
National Yang Ming Chiao Tung University
{tsai.cs09, agarya89.11, chingchun}@nycu.edu.tw hhshuai@nctu.edu.tw

Abstract

Face super-resolution (FSR) is a critical technique for enhancing low-resolution facial images and has significant implications for face-related tasks. However, existing FSR methods are limited by fixed up-sampling scales and sensitivity to input size variations. To address these limitations, this paper introduces an Arbitrary-Resolution and Arbitrary-Scale FSR method with implicit representation networks (ARASFSR), featuring three novel designs. First, ARASFSR employs 2D deep features, local relative coordinates, and up-sampling scale ratios to predict RGB values for each target pixel, allowing super-resolution at any up-sampling scale. Second, a local frequency estimation module captures high-frequency facial texture information to reduce the spectral bias effect. Lastly, a global coordinate modulation module guides FSR to leverage prior facial structure knowledge and achieve resolution adaptation effectively. Quantitative and qualitative evaluations demonstrate the robustness of ARASFSR over existing state-of-the-art methods while super-resolving facial images across various input sizes and up-sampling scales.

1. Introduction

Face super-resolution (FSR) is a technique that focuses on enhancing the resolution and quality of low-resolution facial images. This process involves generating high-resolution (HR) facial images from their low-resolution (LR) counterparts, offering significant advantages in various applications, such as face detection, face recognition, and face parsing. Due to its potential applications, FSR has garnered substantial attention in recent years.

FSR is a specific type of single image super-resolution (SISR) problem. As illustrated in Fig. 1, when addressing the SISR problem, it is assumed that the dataset includes general natural scenes and is not specific to any particular category. Consequently, state-of-the-art (SOTA) SISR methods focus on understanding local image features and learning the super-resolution mapping from LR patches to HR patches. Therefore, patch-based SISR is essentially suitable for handling LR images with varying resolutions. However, facial images possess a more consistent global structure. To leverage this prior knowledge, a more effective FSR method may prefer taking the entire and aligned LR facial image as input and learning the mapping globally to its HR counterpart. However, this global property makes the conventional FSR method sensitive to input image resolution. As a result, the design of FSR necessitates a different approach compared to SISR.

In recent years, deep learning-based methods have made significant strides in improving the performance of FSR. Attention mechanisms have been extensively employed in various FSR methods [3, 13, 24] to enhance facial structural information and texture details. Moreover, some FSR methods leverage face priors [6, 9, 25], such as landmarks and component heatmaps, to further refine the recovery quality. Despite these advances, existing FSR methods exhibit limitations, such as their applicability only to a specific up-sampling scale, which constrains their generalizability. Furthermore, these FSR methods process the entire LR face image as input rather than in patches, making them highly sensitive to variations in input resolution. Consequently, as demonstrated in Fig. 1, changes in input resolution can considerably impact the performance of FSR methods, making them impractical. This implies that when there are changes in up-sampling scales or input resolutions, the existing FSR methods require the network to be retrained.

Facial images in the real world come in various resolutions, presenting challenges for downstream facial-related tasks. To address this issue, it is crucial to develop flexible FSR methods capable of handling a wider range of up-sampling scales and input sizes. For instance, in low-resolution face recognition tasks, some methods [1, 4, 18, 36] first super-resolve LR facial images into HR ones and then perform feature extraction. However, existing FSR methods are limited to a fixed up-sampling scale, while feature extraction networks typically have a fixed input size. As a result, the HR output of FSR must be resized before be-
In this paper, we propose an Arbitrary-Resolution and Arbitrary-Scale FSR method with implicit representation networks (ARASFSR) to address the aforementioned issues. Our method takes 2D deep features, local relative coordinates, and up-sampling scale ratios as inputs of the implicit representation network to predict the RGB value for each target pixel, enabling super-resolution at any up-sampling scale. To further enhance FSR reconstruction, we designed two specialized modules. The first is a local frequency estimation module that predicts high-frequency information about facial texture to reduce the spectral bias effect. The second is a global coordinate modulation module that guides facial structure, allowing for leveraging prior facial knowledge and achieving input resolution adaptation. We conduct comprehensive quantitative and qualitative evaluations, demonstrating our proposed method’s robustness and superiority over SOTA methods. Our main contributions can be summarized as follows:

- We propose ARASFSR, a novel FSR method, with implicit representation networks to address the limitations of FSR in terms of fixed up-sampling scales and sensitivity to input resolution variations.
- The method incorporates two specialized modules to learn high-frequency information and guide facial structure for arbitrary up-sampling scales and variations in input resolutions, resulting in detailed and realistic facial image reconstruction.
- The proposed method outperforms state-of-the-art methods in comprehensive quantitative and qualitative eval-
utions, demonstrating its robustness and superiority in generalized FSR tasks.

2. Related Works

2.1. Face Super-Resolution


2.2. Implicit Neural Representation (INR)

INR is an idea that represents an object using a function that maps coordinates to their corresponding signal values. The success of INR in 3D tasks has led to increasing interest in using INR for 2D image representations. IM-NET [7] introduces an implicit field decoder for 2D shape generation. SIREN [31] leverages sine instead of a ReLU as activation functions for implicit neural representations, resulting in higher image fidelity. LIIF [5] proposes a local implicit image function to learn continuous image representation by taking coordinate and local latent codes as input to predict RGB values at target coordinates, achieving super-resolution at arbitrary up-sampling scale. LTE [21] designs a dominant-frequency estimator for implicit image function to alleviate spectral bias [30] of a standard MLP with ReLU. Finally, DIINN [29] decouples content and positional features using dual interactive implicit neural networks.

However, applying these methods [5, 21, 29] to face images is not straightforward. While coordinates are used to enhance features in LTE [21], they are not used directly as input to MLPs, making it less robust when the up-sampling scale changes. Although LIIF [5] and DIINN [29] take local relative coordinates as MLP’s input, the local implicit image functions still lack a global view, leading to artifacts in the output when the input image size changes. Inspired by the effectiveness of the implicit image function in SISR, ARASFSR extends the flexibility of continuous image representation to face super-resolution. Different from the aforementioned works, ARASFSR has a specific design tailored to face images to alleviate the problem of change in up-sampling scales and be more robust to input size variations.

3. Method

3.1. Overview

Our goal is to develop a general framework for FSR that generates high-quality facial images at arbitrary up-sampling scales while accommodating variations in input resolution. Fig. 2 illustrates the overall architecture of the proposed framework. To achieve arbitrary-scale FSR, we introduce ARASFSR, which utilizes an implicit image function to represent facial images with arbitrary resolution. Besides the modified implicit representation network, our framework integrates two crucial modules: (a) a local frequency estimation module, which captures high-frequency information for facial texture to reduce the spectral bias effect; (b) a global coordinate modulation module, which allows for leveraging prior facial structure knowledge and enabling operation across various input sizes.

3.2. Implicit Image Function

In Local Implicit Image Function approaches [5, 21, 29], the decoding function $f_\theta$ is shared among all images and is parameterized using an MLP with $\theta$ as its model parameters. It maps the latent feature codes and local coordinates to RGB values, i.e., $f_\theta(c, x) : (C, X) \rightarrow S$, where $c \in C$ is a latent feature code from an encoder $E_\phi$ with $\phi$ as its parameters, $x \in X$ is a 2D coordinate in the continuous image domain, and $S$ is a space of predicted RGB values from $f_\theta$.

Our method aims to super-resolve LR facial images $I_{LR} \in \mathbb{R}^{H \times W \times C}$ to HR images $I_{HR} \in \mathbb{R}^{r_y H \times r_x W \times C}$ at any fractional ratio $r_y$ and $r_x$ and arbitrary input resolution $H$ and $W$. Based on the conventional Local Implicit Image Function approaches with modification, for a continuous image $I^{(i)}$, the RGB value $s$ at the target coordinate $x_s$ is defined as follows:

$$s = I^{(i)}(x_s) = f_\theta(c^*, x_s - x_c),$$  \hspace{1cm} (1)

where $c^*$ represents the nearest (Euclidean distance) latent feature code from $x_s$, and $x_c$ denotes the coordinate of the latent code $c^*$ in the image domain.

Feature unfolding. To enrich the feature map, we apply feature unfolding [5, 29], which involves concatenating the features from a $3 \times 3$ neighborhood around each pixel of the feature map. Accordingly, Equation (1) is replaced by:

$$s = I^{(i)}(x_s) = f_\theta(\hat{c}^*, x_s - x_c),$$  \hspace{1cm} (2)

where $\hat{c}^*$ is the enriched local feature code after increasing the number of channels.
Scaling ratio. Although \( f_\theta(c, x) \) can effectively model the continuous distribution, spectral bias tends to make \( f_\theta \) learn a smoother function (i.e., low-frequency function), thereby losing its capacity to recover high-frequency details. To tackle this issue, we propose incorporating an additional scaling ratio as a conditional input to the decoding function. This conditional factor enhances the representational capacity of the original \( f_\theta \). When the scaling ratio is small, \( f_\theta \) is inclined to model smoother functions (less high frequency); conversely, \( f_\theta \) is encouraged to model functions with more high frequency when the scaling ratio is large. Consequently, Equation (2) is extended to:

\[
s = I^i(x_s) = f_\theta(\hat{c}^*, x_s - x_c, r),
\]

where \( r = [1/r_y, 1/r_x] \) represents the scaling ratio.

Local ensemble. The signal prediction at \( x_s \) is done by querying the nearest latent code, and as \( x_s \) moves in the 2D domain, the selection of the nearest latent code can suddenly change, causing the issue of discontinuous prediction. To address this, we refer to [5, 21] and use local ensemble. Accordingly, we modify Equation (3) as follows:

\[
s = I^i(x_s) = \sum_{k \in K} w_k f_\theta(\hat{c}_k^*, x_s - x_{c_k}, r),
\]

where \( K \) is a set of indices of the four nearest latent codes around \( x_s \), \( w_K \) is the bilinear interpolation weight for the enriched latent feature code \( \hat{c}_k \), such that \( \sum_k w_k = 1 \).

3.3. Local Frequency Estimation Module

As previously mentioned, spectral bias is a challenge where a standalone MLP struggles to capture high-frequency textures. To overcome this limitation, we propose a local frequency estimation module. This module employs an encoder-decoder architecture to predict the conditional distribution of high-frequency details \( t \) given an LR facial image \( I^{LR} \), specifically, \( P(t|I^{LR}) \). In particular, the encoder is designed to predict the parameters of a multidimensional Gaussian distribution based on the content of \( I^{LR} \), which models the conditional distribution of the frequency latent code \( P(z|I^{LR}) \). Consequently, for different \( I^{LR} \) instances, the encoder can estimate image-specific high-frequency latent code \( z \) in a probabilistic manner. By sampling a latent code \( z \) from \( P(z|I^{LR}) \) and inputting it into the decoder, we obtain a frequency token \( t \).

After performing feature unfolding, we denote the estimated image-specific high-frequency token of a reference pixel location as \( \hat{t}^* \). By concatenating \( \hat{t}^* \) with the latent feature code \( \hat{c}^* \), we enhance the representation function \( f_\theta \) and update Equation (4) as follows:

\[
s = I^i(x_s) = \sum_{k \in K} w_k f_\theta(\hat{c}_k^*, x_s - x_{c_k}, r),
\]

where \( \hat{p}^* = [\hat{t}^*, \hat{c}^*] \) represents the high-frequency enhanced content feature.

3.4. Global Coordinate Modulation Module

SISR is typically performed using patch-based training, wherein small patches of LR images are used to generate corresponding HR patches. However, in FSR, the entire LR facial image must be mapped to an HR counterpart. Previous SR methods based on INR [5, 21, 29] suffer from artifacts when the input resolution of facial images changes because the implicit image function only possesses a local view. In contrast, FSR methods process the entire LR facial image as input, making the global coordinate an essential component that provides landmark locations for re-
constructing facial details. Furthermore, the global coordinate is resilient to variations in up-sampling scales and input sizes, making it an ideal source of facial priors.

Intuitively, if facial images are pre-aligned, a well-trained model may be able to predict the possible texture around a specific location based on the normalized coordinate and facial structure. To provide the implicit image function with a global perspective, we propose a global coordinate guidance module that incorporates the global coordinate \( x_s \) as input. In our implementation, we employ positional encoding \([28, 34]\), as illustrated in Equation (6), to map the coordinates to a higher dimension before inputting them into the MLP. Note that each of the two coordinate values in \( x_s \) is normalized to lie within the range \([-1, 1]\) before positional encoding. In our experiments, we set \( N = 10 \).

\[
g_s = (x_s, \sin(2^0 \pi x_s), \cos(2^0 \pi x_s), \ldots, \\
\sin(2^{N-1} \pi x_s), \cos(2^{N-1} \pi x_s)).
\] (6)

Recent approaches have addressed the spectral bias issue by employing non-linear activations. SIREN \([31]\) utilizes the sine layer, which leads to rapid convergence and high data fidelity. Additionally, works such as \([27, 29]\) adopted modulated periodic activations. Inspired by them, we implement periodic activations to predict the modulation parameters given \( g_s \). Next, as illustrated in Fig. 2, we modulate the local implicit function \( f_0 \) by multiplying and concatenating in order to fuse the original feature of the MLP with the encoded global coordinate information. The process is detailed in the following equations:

\[
s_0 = \text{ReLU}(w_0 \left[ \hat{g}_k, x_s - x_{c_k}, r \right] + b_0). \tag{7}
\]

\[
g_0 = \sin(w_0 g_s + b'_0). \tag{8}
\]

\[
m_0 = s_0 \odot \hat{g}_0. \tag{9}
\]

\[
s_i = \text{ReLU}(w_i \left[ m_{i-1}, s_{i-1} \right] + b_i). \tag{10}
\]

\[
g_i = \sin(w'_i g_s + b'_i). \tag{11}
\]

\[
m_i = s_i \odot \hat{g}_i. \tag{12}
\]

In the above equations, \( w_i \) and \( w'_i \) are the weights, \( b_i \) and \( b'_i \) are the biases, and \( s_i \) is the latent feature of the \( i^{th} \) layer within the MLP. \( \hat{g}_i \) is the global positional feature and \( m_i \) is the modulated output. The last output of \( s_i \) is then passed through a final dense layer to output the predicted RGB value. With the guidance of the global coordinate, the modulated implicit image function now has a global view to identify the potential location of the landmark.

In summary, the global coordinate modulation module incorporates prior facial structure information into FSR by mapping the global coordinate into a high-dimensional space using position encoding. By combining global coordinate guidance with locally enhanced features, our ARAS-FSR achieves improved facial reconstruction results.

### 3.5. Skip Connection

The effectiveness of long skip connections in learning high-frequency components and improving convergence stability has been demonstrated in residual networks \([16, 21]\). In our proposed architecture, we introduce an additional branch incorporating a skip connection. Including this long skip connection can alleviate information loss and enhance the network’s capability to capture fine details.

### 3.6. Loss Function

We train our model end-to-end and use the Charbonnier loss \([19]\) given by:

\[
L_\delta = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(I^{HR}(x_s) - I^{SR}(x_s))^2 + \delta^2}, \tag{13}
\]

where \( I^{HR}(x_s) \) is the ground truth value, \( I^{SR}(x_s) \) is the predicted value, \( N \) is the total number of samples, and \( \delta \) is a hyper-parameter that controls the smoothness of the loss function.

### 4. Experiments

#### 4.1. Datasets and Metrics

We carry out experiments on four datasets. Firstly, we compare the performance of our approach with INR-based SISR methods at different up-sampling scales on CelebAHQ dataset \([15]\). This dataset consists of 30,000 HR face images (1024x1024 pixels) selected from CelebA dataset \([23]\). Secondly, we create CelebAHQ-NN-JPEG dataset by down-sampling CelebAHQ dataset using the nearest-neighbor method and introducing JPEG compression artifacts. This dataset, along with SCFace dataset \([11]\) containing facial images captured by surveillance cameras, demonstrates our method’s applicability in real-world scenarios. Lastly, we compare our approach with FSR methods on CelebAHQ and Helen datasets \([20]\). We evaluate the performance using two widely-used metrics, PSNR and SSIM \([33]\), calculated on the Y channel in the YCbCr color space.

#### 4.2. Implementation Details

Since CelebAHQ is a high-quality dataset with pre-aligned faces, no further alignment is necessary. For Helen and SCFace, we align and crop all facial images with respect to their landmarks using MTCNN \([37]\). In FSR, an LR face image of size \( L_r \times L_r \) is mapped to its HR counterpart of size \( H_r \times H_r \). To evaluate the effectiveness of different up-sampling scales, multiple LR-HR pairs are required. To achieve this, the ground-truth images \( D^{GT} \) (e.g., 1024x1024 high-quality images in CelebAHQ) are down-sampled using bicubic interpolation to produce the
Table 1. Quantitative comparison on CelebAHQ with INR-based SISR methods (PSNR(dB)). The best and second best performances are highlighted in red and blue colors, respectively. "E-" and "R-" indicate the use of EDSR and RDN as encoders, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>In-distribution</th>
<th>Out-of-distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>×1.5 ×2</td>
<td>×4 ×8</td>
</tr>
<tr>
<td>E-DIINN [29]</td>
<td>40.1781 37.0943</td>
<td>32.4799 30.4560</td>
</tr>
<tr>
<td>E-ARASFSR</td>
<td>40.1954 37.1097</td>
<td>32.4799 30.9148</td>
</tr>
<tr>
<td>R-DIINN [29]</td>
<td>40.3033 37.1676</td>
<td>32.0622 30.6020</td>
</tr>
<tr>
<td>R-ARASFSR</td>
<td>40.3100 37.1830</td>
<td>32.4987 30.9741</td>
</tr>
</tbody>
</table>

corresponding HR images \(D^{HR}\) at variant target HR levels \(\{H_r\}\). Similarly, we attain LR images \(D^{LR}\) at different target LR levels \(\{L_r\}\) by using either bicubic or nearest-neighbor down-sampling with augmented JPEG compression noise. However, to validate model generalization across input resolution, we fix the target \(L_r\) level and vary the target \(H_r\) level according to up-sampling scales during training. Furthermore, we use Charbonnier loss [19] and Adam optimizer [17]. The models are trained for 200 epochs with a batch size of 16. The initial learning rate is 1e-4 and is reduced by a factor of 0.1 at epoch 100.

4.3. Comparison with State-of-the-Arts

4.3.1 Comparison with INR-based SISR methods

We compare our proposed method with SOTA INR-based SISR methods such as LIIF [5], LTE [21], and DIINN [29], on different scenarios.

Evaluation on CelebAHQ: To evaluate the performance of various up-sampling scales with INR-based SISR methods, we assess up-sampling tasks involving both in-distribution and out-of-distribution scales. Considering that the average image resolution of large face datasets, such as Vggface2 [2], is typically below 180×180 pixels without background cropping, we uniformly sample up-sampling scales within \(\{\times1 \sim \times2\}\) and set \(L_r = 64\) and \(H_r \in \{64 \sim 128\}\) during training. During testing, we evaluate our method with much larger up-sampling scales within \(\{\times1.5 \sim \times8\}\). Table 1 presents a quantitative comparison on CelebAHQ. The top and bottom rows show the results when EDSR [22] and RDN [38] are used as encoders, respectively. ARASFSR achieves comparable performance for in-distribution scales and outperforms other methods for out-of-distribution scales. This demonstrates the superior generalizability of ARASFSR to arbitrary precision, which is crucial in real-world applications where the desired up-scaling factor may not be known beforehand. Furthermore, Fig. 3 presents a visual comparison using RDN as the encoder. As the up-sampling scale increases, we find that INR-based SISR methods produce SR results with blocking artifacts when zooming into the SR images. In contrast, ARASFSR generates stable results with plausible details.

Real-world case on CelebAHQ-NN-JPEG: To evaluate the effectiveness in real-world scenarios, we assess the performance of INR-based SISR methods on CelebAHQ-NN-JPEG, synthesized by performing nearest neighbor down-sampling and augmenting with JPEG compression artifacts to mimic surveillance faces. We train the up-sampling scales within \(\{\times1 \sim \times2\}\) with a resolution setting of \(L_r = 32\) and \(H_r \in \{32 \sim 64\}\). To compare the performance of different input LR resolutions during testing, we evaluate the models on the following \(\times2\) super-resolution scenarios, using different input resolutions of 64(LR)-128(HR). Fig. 4 offers a visual comparison with EDSR employed as the encoder. Previous works have encountered artifacts when the input resolution of facial images changes, as the implicit image function is constrained to a local view. The arrows indicate misplaced landmarks (such as smaller eyes) and artifacts (such as distortions in the nose). In contrast, our results display clear facial landmarks.

Real-world case on SCface: To further support the effectiveness in surveillance scenarios, we evaluate our approach using the d1 subset of SCface. This subset contains images captured by surveillance cameras from a distance of 4.2m, which is the most challenging subset in SCface. The visual comparison in Fig. 5 shows the SR results of LR probes using different methods. Their HR gallery images are pro-
vided for comparison. It is evident from the comparison that existing methods encounter challenges in accurately preserving and reconstructing facial features, especially the eyes. In contrast, our method demonstrates exceptional performance in retaining and restoring fine facial details.

4.3.2 Comparison with FSR methods

We present a comparative analysis of our proposed method with SOTA FSR methods such as FSRNet [6], PFSRNet [9], DICNet [25], SPARNet [3], SISN [24] and MRRNet [13]. In our comparison with FSR methods, we utilize RDN as a encoder. Since FSR methods focus on fixed-scale up-sampling, we train the networks for $\times 8$ super-resolution with $L_r = 16$ and $H_r = 128$. However, during testing, we evaluate not only the match scenario (i.e., 16(LR)-128(HR)) but also the mismatch cases (i.e, 12(LR)-96(HR) and 64(LR)-128(HR)). As there is no standardized benchmark for comparing FSR methods, we take the following approach to ensure a fair comparison. For methods that provide training codes, we retrain these models using our training set on CelebAHQ and Helen. For methods that do not provide training codes, we test these models using their pretrained weights on CelebA in our CelebAHQ experiments. However, we do not evaluate the performance of these models on Helen since their training codes or pretrained weights are not available.

Evaluation on CelebAHQ: We first conduct experiments on CelebAHQ. The quantitative results presented in Table 2 indicate that our method produces comparable performance to other FSR methods in both the match and mismatch scenarios. The visual comparison shown in Fig. 6 illustrates that our method produces high-quality results with minimal artifacts and clearer facial features. Unlike existing FSR methods that are only suitable for fixed up-sampling scales and input resolutions, our method can super-resolve images at any desired scale and is robust to input resolution, making it more versatile.

Evaluation on Helen: We conducted additional experiments on Helen in the match scenario (i.e., 16(LR)-128(HR)). The quantitative results in Table 2 demonstrate that our method achieves results comparable to those of other FSR methods. Notably, our method can also be used with a stronger encoder to achieve even better results.

4.4. Ablation Study

ARASFSR comprises implicit representation networks with three primary modules. We retrain the model with EDSR for three variants: ARASFSR without the local frequency estimation module (-L), ARASFSR without the global coordinate modulation module (-G), and ARASFSR without a skip connection (-S). To assess the effectiveness of each module, we conduct a quantitative ablation study, with the results presented in Table 3. Additionally, a qualitative ablation study depicted in Fig. 7 examines the super-resolved results and the difference maps between the bicubic up-sampled results and the super-resolved results.

Our findings reveal that the local frequency estimation module plays a crucial role in capturing frequency information for facial texture, as evidenced by the comparison between ARASFSR and ARASFSR(-L). Moreover, the use of global coordinate modulation provides valuable guidance for fusing local features and global facial landmark priors, as demonstrated by the contrast between ARASFSR and ARASFSR(-G). Finally, we observe that the incorporation of a skip connection consistently enhances the quality of ARASFSR compared to ARASFSR(-S).

5. Conclusion

In this paper, we proposed an Arbitrary-Resolution and Arbitrary-Scale FSR method. Our framework employs an implicit image function that can effectively handle changes in the up-sampling scale. Furthermore, our approach incorporates a local frequency estimation module to capture high-frequency information for facial texture and a
Table 2. Quantitative comparison on CelebAHQ and Helen with FSR methods. The best and second best performances are highlighted in red and blue colors, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>CelebAHQ</th>
<th>Helen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mismatch match mismatch match</td>
<td>match match</td>
</tr>
<tr>
<td></td>
<td>×8</td>
<td>×8</td>
</tr>
<tr>
<td></td>
<td>12-96</td>
<td>16-128</td>
</tr>
<tr>
<td>PSNR</td>
<td>SSIM</td>
<td>PSNR</td>
</tr>
<tr>
<td>FSRNet [6]</td>
<td>22.4421 0.6562 24.4323 0.7512</td>
<td>25.1409 0.7479</td>
</tr>
<tr>
<td>PFSRNet [9]</td>
<td>22.6620 0.6440 20.0664 0.5545</td>
<td>22.2394 0.6108</td>
</tr>
<tr>
<td>DICNet [25]</td>
<td>22.4446 0.6112 26.4781 0.7758</td>
<td>27.2324 0.7631</td>
</tr>
<tr>
<td>SPARNet [3]</td>
<td>24.5753 0.7086 26.5321 0.7792</td>
<td>28.2184 0.7907</td>
</tr>
<tr>
<td>SISN [24]</td>
<td>24.7083 0.7166 26.6575 0.7807</td>
<td>28.3367 0.7946</td>
</tr>
<tr>
<td>MRRNet [13]</td>
<td>24.7207 0.7192 26.5140 0.7755</td>
<td>28.3538 0.7943</td>
</tr>
<tr>
<td>ARASFSR</td>
<td>24.7712 0.7214 26.6400 0.7813</td>
<td>28.3088 0.7948</td>
</tr>
</tbody>
</table>

Figure 6. Visual comparison on CelebAHQ with FSR methods.

Table 3. Quantitative Ablation study of ARASFSR on CelebAHQ (PSNR(dB)).

<table>
<thead>
<tr>
<th>Method</th>
<th>In-distribution</th>
<th>Out-of-distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>×1.5 ×2</td>
<td>×4 ×8</td>
</tr>
<tr>
<td></td>
<td>64-96 64-128</td>
<td>64-256 64-512</td>
</tr>
<tr>
<td>ARASFSR</td>
<td>40.1954 37.1097</td>
<td>32.4799 30.9148</td>
</tr>
<tr>
<td>ARASFSR(-L)</td>
<td>40.1827 37.0936</td>
<td>32.3301 30.6935</td>
</tr>
<tr>
<td>ARASFSR(-G)</td>
<td>40.1666 37.0848</td>
<td>32.3210 30.6504</td>
</tr>
<tr>
<td>ARASFSR(-S)</td>
<td>40.1336 36.9969</td>
<td>32.3958 30.8488</td>
</tr>
</tbody>
</table>

global coordinate modulation module to guide the facial structure, ensuring effective operation across various input sizes. Through quantitative and qualitative experiments, we demonstrated the robustness and superiority of our proposed method compared to state-of-the-art methods.

Acknowledgements. Thanks to Yi-ren Ye and Hsuan-Tung Liu for valuable discussions of this work. This work was funded in part by E-SUN COMMERCIAL BANK, LTD. and also supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC-112-2221-E-A49-089-MY3, Grant NSTC-110-2221-E-A49-066-MY3, Grant NSTC-111-2634-F-A49-010, Grant NSTC-112-2425-H-A49-001, and in part by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and the Ministry of Education (MOE), Taiwan.
References


