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-

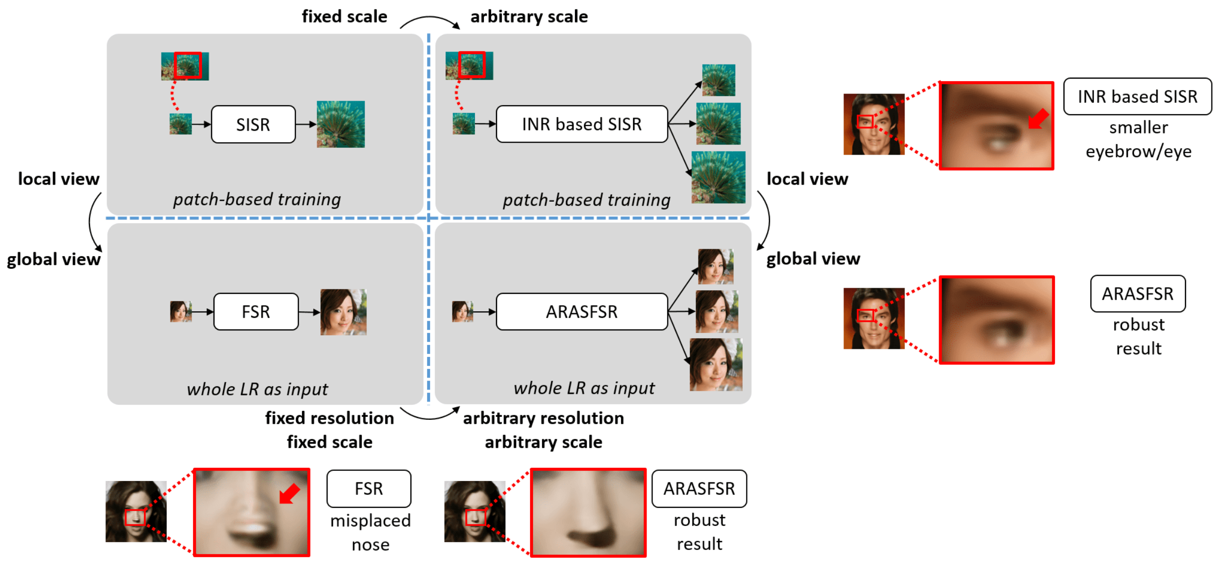

Figure 1. Comparison of our proposed arbitrary-resolution and arbitrary-scale face super-resolution (ARASFSR) method with single image super-resolution (SISR) methods, face super-resolution (FSR) methods, and implicit neural representation (INR) based SISR methods. The figure highlights the differences in patch-based training in SISR versus the need for the whole low-resolution face image as input to match the test-time input distribution in FSR, resulting in a global approach instead of SISR's local view. Conventional FSR methods have limited practical applications due to their fixed up-sampling scales and sensitivity to input resolution variations, while applying INR-based SISR methods to face images is not straightforward due to the lack of a global view. To address these issues, we propose ARASFSR, which utilizes implicit representation networks to produce facial images of any resolution and can handle changes in input resolution.

ing fed into the feature extraction network, which compromises data fidelity and may negatively impact the downstream task. Therefore, there is an urgent need to develop a generalized FSR method that can handle a broader range of up-sampling scales and input sizes, making FSR more suitable for various facial-related applications.

Implicit neural representation (INR) has gained popularity due to its effectiveness in 3D shape reconstruction, as demonstrated in recent works such as [8, 10, 14, 28, 32]. The fundamental concept behind INR involves representing objects as multi-layer perceptrons (MLPs) that map coordinates to corresponding signal values. In recent years, there has been growing interest in using INR to learn 2D continuous image representations [5, 21, 26, 29, 34]. This approach utilizes implicit image functions, which are MLPs that input local coordinate information and query local latent codes to predict RGB values at target coordinates. This coordinates continuity property enables the representation of images in arbitrary resolutions, making it possible to achieve super-resolution at any up-sampling scale. However, existing methods based on local implicit image functions are not directly applicable to FSR due to their lack of a global view and designs that incorporate facial prior knowledge. Furthermore, as illustrated in Fig. 1, these methods are sensitive to input image resolution, similar to existing FSR methods.

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-

uations, demonstrating its robustness and superiority in generalized FSR tasks.

2. Related Works

2.1. Face Super-Resolution

Deep learning-based methods have made significant strides in the field of FSR. URDGN [35] uses a deep discriminative generative network to super-resolve ultra-LR facial images. Wavelet-SRNet [12] employs a wavelet-based CNN method to predict the corresponding series of HR wavelet coefficients from an LR input image. Some methods incorporate facial priors in the super-resolution process. FSRNet [6] utilizes facial landmark heatmaps and parsing maps. PFSRNet [9] proposes a facial attention loss by multiplying the pixel difference and heatmap values to progressively recover facial details. DIC [25] adopts an iterative collaboration based on the guidance of landmark maps. Attention mechanisms have also been adopted in the latest methods. SPARNet [3] makes use of a spatial attention mechanism to adaptively guide features associated with facial structures. SISN [24] introduces a split-attention in the split-attention network to strengthen facial structure and details. MRRNet [13] designs a spatial attention mechanism guided by multi-scale receptive field features.

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_θ is shared among all images and is parameterized using an MLP with θ as its model parameters. It maps the latent feature codes and local coordinates to RGB values, i.e., $f_\theta(c, x) : (\mathcal{C}, \mathcal{X}) \mapsto \mathcal{S}$, where $c \in \mathcal{C}$ is a latent feature code from an encoder E_φ with φ as its parameters, $x \in \mathcal{X}$ is a 2D coordinate in the continuous image domain, and $\mathcal{S}$ is a space of predicted RGB values from f_θ.

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:

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

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 3×3 neighborhood around each pixel of the feature map. Accordingly, Equation (1) is replaced by:

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

where $\hat{c}^*$ is the enriched local feature code after increasing the number of channels.

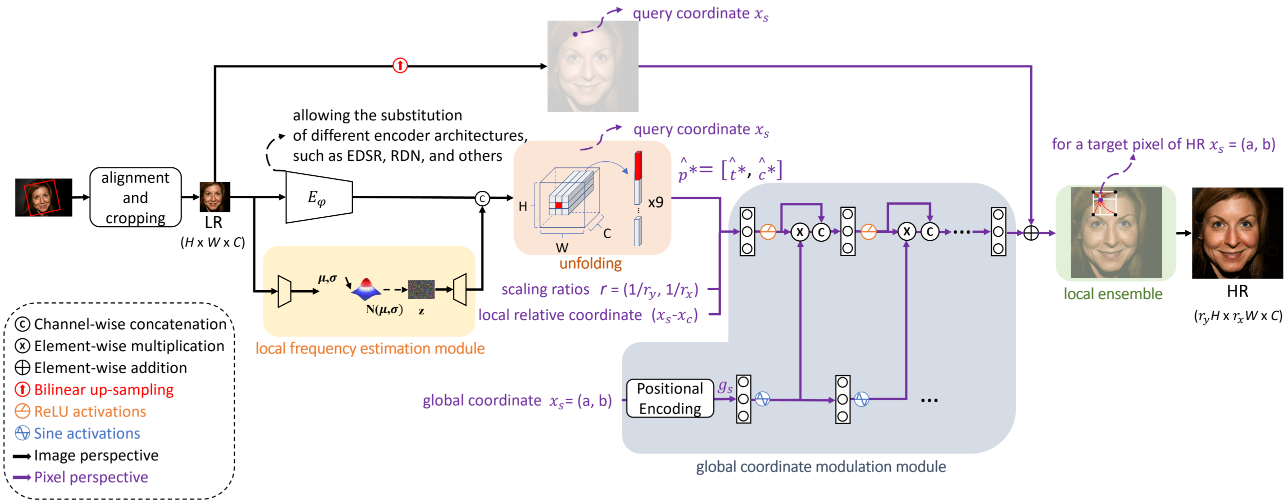

Figure 2. Overall architecture of the proposed framework.

Scaling ratio. Although $f_\theta(c, x)$ can effectively model the continuous distribution, spectral bias tends to make f_θ 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_θ. When the scaling ratio is small, f_θ is inclined to model smoother functions (less high frequency); conversely, f_θ is encouraged to model functions with more high frequency when the scaling ratio is large. Consequently, Equation (2) is extended to:

$$s = I^{(i)}(x_s) = f_\theta(\hat{c}^*, x_s - x_c, r), \tag{3}$$

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:

$$s = I^{(i)}(x_s) = \sum_{k \in K} w_k f_\theta(\hat{c}^*_k, x_s - x_{c_k}, r), \tag{4}$$

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 multi-dimensional 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_θ and update Equation (4) as follows:

$$s = I^{(i)}(x_s) = \sum_{k \in K} w_k f_\theta(\hat{p}^*_k, x_s - x_{c_k}, r), \tag{5}$$

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 reconstructing 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 $N = 10$.

$$\begin{aligned} g_s = (x_s, sin(2^0\pi x_s), cos(2^0\pi x_s), ..., \\ sin(2^{N-1}\pi x_s), cos(2^{N-1}\pi x_s)). \end{aligned} \tag{6}$$

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_θ 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:

$$s_0 = ReLU(w_0\,[\hat{p}_k^*, x_s - x_{c_k}, r] + b_0). \tag{7}$$

$$\hat{g}_0 = sin(w_0' g_s + b_0'). \tag{8}$$

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

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

$$\hat{g}_i = sin(w_i' g_s + b_i'). \tag{11}$$

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

In the above equations, w_i and w_i' are the weights, b_i and b_i' are the biases, and s 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 δ 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 (1024×1024 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×L_r is mapped to its HR counterpart of size H_r×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., 1024×1024 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.

Method	In-distribution ×1.5 64-96	In-distribution ×2 64-128	Out-of-distribution ×4 64-256	Out-of-distribution ×8 64-512
E-LIIF [5]	40.1002	36.9650	32.3501	30.8161
E-LTE [21]	40.1913	37.1062	31.2708	28.4518
E-DIINN [29]	40.1781	37.0943	32.1410	30.4560
E-ARASFSR	40.1954	37.1097	32.4799	30.9148
R-LIIF [5]	40.2433	37.1188	32.3950	30.7763
R-LTE [21]	40.2813	37.2104	31.9483	29.9873
R-DIINN [29]	40.3033	37.1676	32.0622	30.6020
R-ARASFSR	40.3100	37.1830	32.4987	30.9741

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 $\{\times 1 \sim \times 2\}$ 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 $\{\times 1.5 \sim \times 8\}$. 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-

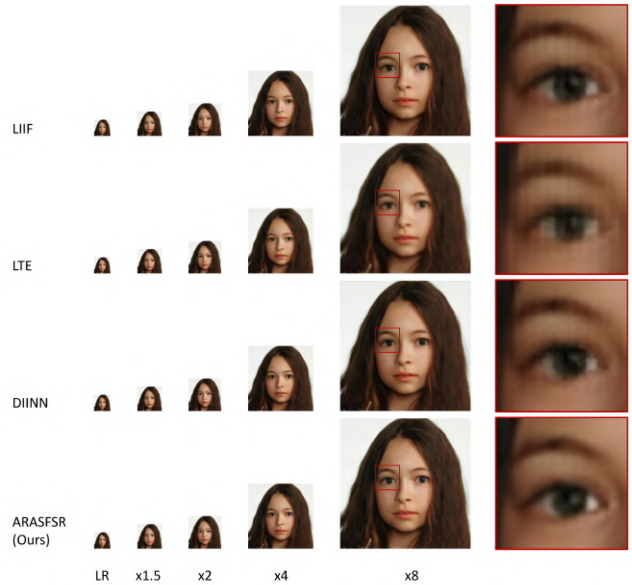

Figure 3. Visual comparison on CelebAHQ with INR-based SISR methods.

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 $\{\times 1 \sim \times 2\}$ 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 $\times 2$ 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-

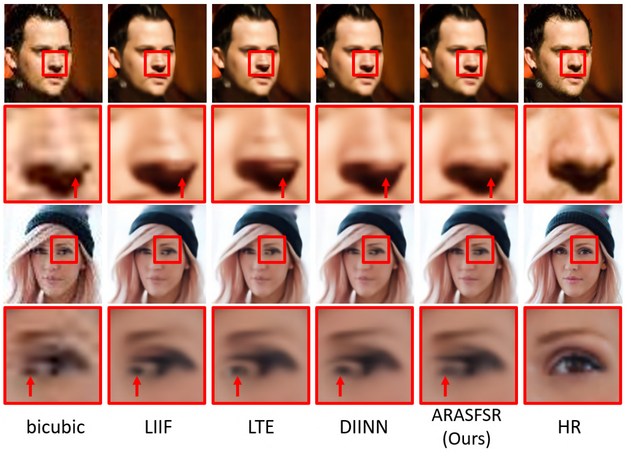

Figure 4. Visual comparison of real world case on CelebAHQ-NN-JPEG with INR-based SISR methods.

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

Figure 5. Visual comparison of real-world case on SCface with INR-based SISR methods. Please zoom in for details.

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.

Method	CelebAHQ						Helen	
	mismatch ×8 12-96		match ×8 16-128		mismatch ×8 32-256		match ×8 16-128	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FSRNet [6]	22.4421	0.6562	24.4323	0.7512	25.1409	0.7479	-	-
PFSRNet [9]	22.6620	0.6440	20.0664	0.5545	22.2394	0.6108	-	-
DICNet [25]	22.4446	0.6112	26.4781	0.7758	27.2324	0.7631	25.1722	0.7265
SPARNet [3]	24.5753	0.7086	26.5321	0.7792	28.2184	0.7907	25.5450	0.7509
SISN [24]	24.7083	0.7166	26.6575	0.7807	28.3367	0.7946	24.9867	0.7160
MRRNet [13]	24.7207	0.7192	26.5140	0.7755	28.3538	0.7943	25.1263	0.7260
ARASFSR	24.7712	0.7214	26.6400	0.7813	28.3088	0.7948	25.6002	0.7445

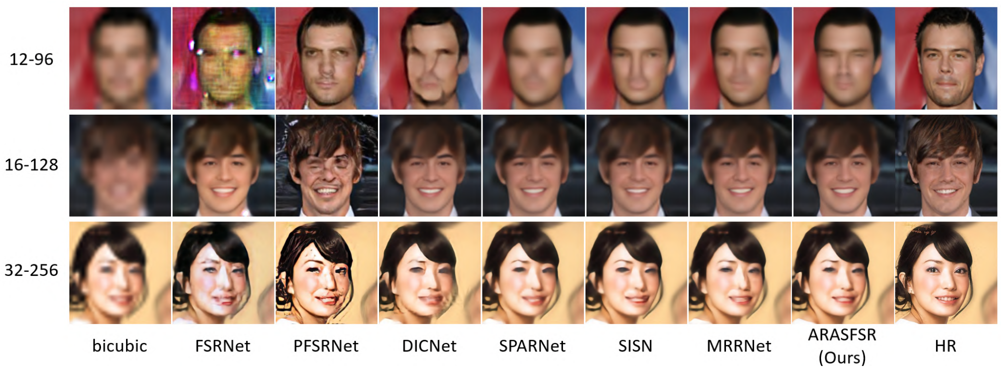

Figure 6. Visual comparison on CelebAHQ with FSR methods.

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

Method	In-distribution		Out-of-distribution	
	×1.5 64-96	×2 64-128	×4 64-256	×8 64-512
ARASFSR	*40.1954*	*37.1097*	*32.4799*	*30.9148*
ARASFSR(-L)	40.1827	37.0936	32.3301	30.6935
ARASFSR(-G)	40.1666	37.0848	32.3210	30.6504
ARASFSR(-S)	40.1336	36.9969	32.3958	30.8488

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.

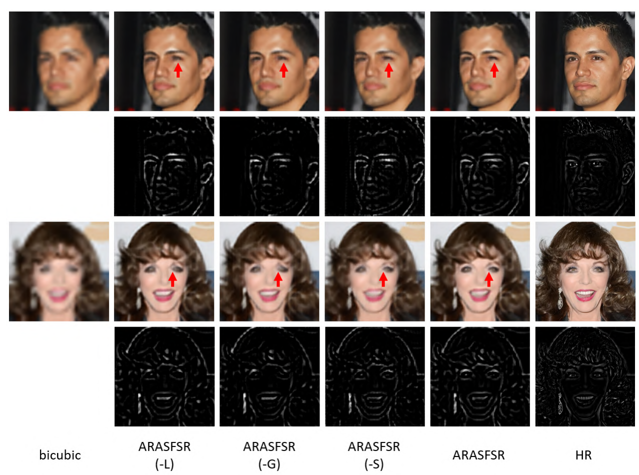

Figure 7. Qualitative ablation study of ARASFSR on CelebAHQ (LR-HR: 32-256). For a better view of the details, please zoom in.

References

[1] Qiqi Bao, Rui Zhu, Bowen Gang, Pengyang Zhao, Wenming Yang, and Qingmin Liao. Distilling resolution-robust identity knowledge for texture-enhanced face hallucination. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6727–6736, 2022.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[3] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K. Wong. Learning spatial attention for face super-resolution. 2020.

[4] Jin Chen, Jun Chen, Zheng Wang, Chao Liang, and Chia-Wen Lin. Identity-aware face super-resolution for low-resolution face recognition. *IEEE Signal Processing Letters*, 27:645–649, 2020.

[5] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021.

[6] 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*, 2018.

[7] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[9] Kim Deokyun, Kim Minseon, Kwon Gihyun, and Kim Dae-Shik. Progressive face super-resolution via attention to facial landmark. In *Proceedings of the 30th British Machine Vision Conference (BMVC)*, 2019.

[10] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[12] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1689–1697, 2017.

[13] Weikang Huang, Shiyong Lan, Wenwu Wang, Xuedong Yuan, Hongyu Yang, Piaoyang Li, and Wei Ma. Face super-resolution with spatial attention guided by multiscale receptive-field features. In *International Conference on Artificial Neural Networks,(ICANN2022)*, 2022.

[14] Chiyu Max Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

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

[18] Shun-Cheung Lai, Chen-Hang He, and Kin-Man Lam. Low-resolution face recognition based on identity-preserved face hallucination. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1173–1177. IEEE, 2019.

[19] 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.

[20] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pages 679–692. Springer, 2012.

[21] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1929–1938, June 2022.

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[24] Tao Lu, Yuanzhi Wang, Yanduo Zhang, Yu Wang, Liu Wei, Zhongyuan Wang, and Junjun Jiang. Face hallucination via split-attention in split-attention network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5501–5509, 2021.

[25] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[26] Cheng Ma, Peiqi Yu, Jiwen Lu, and Jie Zhou. Recovering realistic details for magnification-arbitrary image super-resolution. *IEEE Transactions on Image Processing*, 31:3669–3683, 2022.

[27] Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, and Manmohan Chandraker. Modulated periodic activations for generalizable local functional representations. In *Proceedings of the IEEE/CVF Interna-*

tional Conference on Computer Vision, pages 14214–14223, 2021.

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[29] Quan H. Nguyen and William J Beksi. Single image super-resolution via a dual interactive implicit neural network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4936–4945, 2023.

[30] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[31] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[32] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[34] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021.

[35] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 318–333. Springer, 2016.

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

[37] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[38] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.