Arbitrary-Resolution and Arbitrary-Scale Face Super-Resolution with Implicit Representation Networks: Supplementary Material

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

1. Coordinate for a Continuous Image

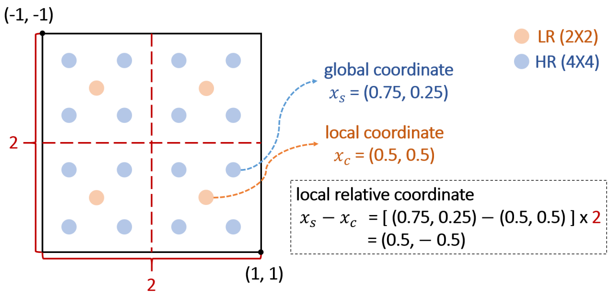

Figure 1. An example of calculating local, global, and local relative coordinates.

For a continuous image $I^{(i)}$, we define the global coordinates as the coordinates that reference the position of a pixel with respect to the center of the image (i.e., the center is the origin of the global coordinate system). For a given target global coordinate x_s corresponding to the HR image, we refer to the coordinate of the nearest LR pixel x_c as the origin of the local coordinate system. The local relative coordinate, which serves as the input to the implicit image function, is then obtained by subtracting x_c from x_s (i.e., x_s-x_c). Both the global coordinate and the local relative coordinate are then scaled to the interval $[-1, 1]$. It is worth noting that these coordinate calculations remain invariant regardless of the overall image size. To illustrate this concept, consider the example presented in Fig. 1, which involves a pair of LR (2×2)-HR (4×4) images.

For a target pixel with global coordinate $x_s = (0.75, 0.25)$ in the HR image, the corresponding local coordinate system has $x_c = (0.5, 0.5)$ as the origin. By computing $x_s - x_c = (0.25, -0.25) \times 2$, we obtain the local relative coordinate as (0.5, -0.5). The reason for multiplying by 2 is to ensure that the local relative coordinate is within the range $[-1, 1]$, providing a normalized representation irrespective of the grid dimensions.

2. Architecture Details

We utilize either EDSR [4] or RDN [7] as encoders, excluding their up-sampling layers, for baseline feature extraction. The encoder generates a latent code with the same width and height as the input LR image. Subsequently, we concatenate the extracted latent code with the output frequency token from the proposed local frequency estimation module and perform the unfolding operations to enhance the extracted feature map. The decoder in our proposed method, referred to as the global coordinate modulation module, is composed of two 4-layer MLPs. Each MLP consists of 256 hidden units and utilizes ReLU and sine activations. Following these MLPs, a final dense layer with 256 hidden units is employed. During the decoder stage, we concatenate the 2D deep features, local relative coordinates, and up-sampling scale ratios, which serve as the input to one MLP. Additionally, we modulate this input with the encoded global coordinate information from the other MLP. This combination enables us to incorporate global coordinate guidance and predict RGB values for each target pixel using the final dense layer. By querying every pixel, we can produce the entire image according to the scaling ratios.

3. Dataset Split Details

For CelebAHQ and CelebAHQ-NN-JPEG datasets, we adopt the same dataset split as previous works [1], utilizing 25,000 images for training and 5,000 images for testing. In the case of Helen dataset, we follow the same split as previous works [2, 5], using 2,000 images for training and 50 images for testing.

4. Additional Comparison Results

We have provided a detailed comparison between our proposed method and state-of-the-art (SOTA) INR-based SISR methods across a variety of scenarios. Specifically, we evaluated the performance of our approach under different upsampling scales and compared its effectiveness with other methods when different input LR resolutions were used dur-

Table 1. Quantitative comparison on CelebAHQ with INR-based SISR methods: exploring various input resolutions and up-sampling scales (PSNR (dB)). The best and second best performances are highlighted in red and blue colors, respectively.

Method	smaller input resolution			larger input resolution	
	×4	×8	×16	×2.6	×5.3
	32-128	32-256	32-512	96-256	96-512
LIIF [1]	29.7036	27.8170	27.0594	35.2839	33.1212
LTE [3]	28.6614	25.6519	24.7687	35.0835	31.3316
DIINN [6]	29.5231	27.5411	26.7018	35.2048	32.8328
ARASFSR	29.8574	27.9548	27.1823	35.3317	33.1916

Table 2. Quantitative comparison of real-world case on CelebAHQ-NN-JPEG with INR-based SISR methods: exploring various input resolutions and up-sampling scales (PSNR (dB)). The best and second best performances are highlighted in red and blue colors, respectively.

Method	smaller input resolution			larger input resolution	
	×4	×8	×16	×2.6	×5.3
	16-64	16-128	16-256	48-128	48-256
LIIF [1]	22.7435	22.0485	21.6715	28.4689	27.1927
LTE [3]	22.6693	21.3579	20.6498	28.5183	26.7715
DIINN [6]	22.8666	22.1856	21.8145	28.5524	27.2923
ARASFSR	22.9398	22.2392	21.8591	28.6026	27.3158

ing testing. As an extension to this analysis, below, we further evaluate our ASARFSR method in comparison with INR-based SISR methods under conditions that *simultaneously present different upsampling scales and varied input resolutions*. For these experiments, we employ EDSR as the encoder.

4.1. Evaluation on CelebAHQ

During the training phase, we employ a uniform sampling strategy to select up-sampling scales from the range of $\{\times 1 \sim \times 2\}$, utilizing a resolution setting of $L_r = 64$ and $H_r \in \{64 \sim 128\}$. Subsequently, during the testing phase, we evaluate our method's performance on smaller and larger input LR resolutions, encompassing out-of-distribution scales that exceed $\times 2$.

Table 1 presents a quantitative comparison on CelebAHQ. The proposed method, ARASFSR, outperforms other approaches across smaller and larger input resolutions, demonstrating its exceptional generalizability in varying out-of-distribution scales.

Furthermore, Fig. 2 visually compares INR-based SISR methods specifically for the LR resolution of 32 and the HR resolution of 256. By closely examining the zoom-in regions, it becomes evident that ARASFSR generates more precise details, while INR-based SISR methods tend to produce SR results that exhibit noticeable blocking and blurry artifacts.

4.2. Real-world cases on CelebAHQ-NN-JPEG

To evaluate the effectiveness of INR-based SISR methods in real-world scenarios, we conduct performance assessments on CelebAHQ-NN-JPEG. During training, we train the up-sampling scales within the range of $\{\times 1 \sim \times 2\}$, utilizing a resolution setting of $L_r = 32$ and $H_r \in \{32 \sim 64\}$. Subsequently, during testing, we evaluate the performance of our method on both smaller and larger input LR resolutions, including out-of-distribution scales that exceed $\times 2$.

The quantitative comparison of different methods is presented in Table 2. Our proposed method demonstrates superior performance compared to other methods and exhibits robustness when compared to other INR-based SISR methods.

Additionally, Fig. 3 provides a visual comparison of INR-based SISR methods under the 48(LR)-256(HR) testing scenario. It is evident that other methods encounter artifacts when the input LR resolution changes, primarily due to the lack of a global view in the implicit image function. As a consequence, this can result in misplaced or incorrectly sized facial landmarks, such as eyes, nose, and mouth, in the super-resolved output. The red arrows indicate an example of such artifacts. In contrast, our proposed method effectively restores facial details without introducing these artifacts.

References

[1] 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. 1, 2

[2] Weikang Huang, Shiyong Lan, Wenwu Wang, Xuedong Yuan, Hongyu Yang, Piaoyang Li, and Wei Ma. Face super-resolution with spatial attention guided by multiscale

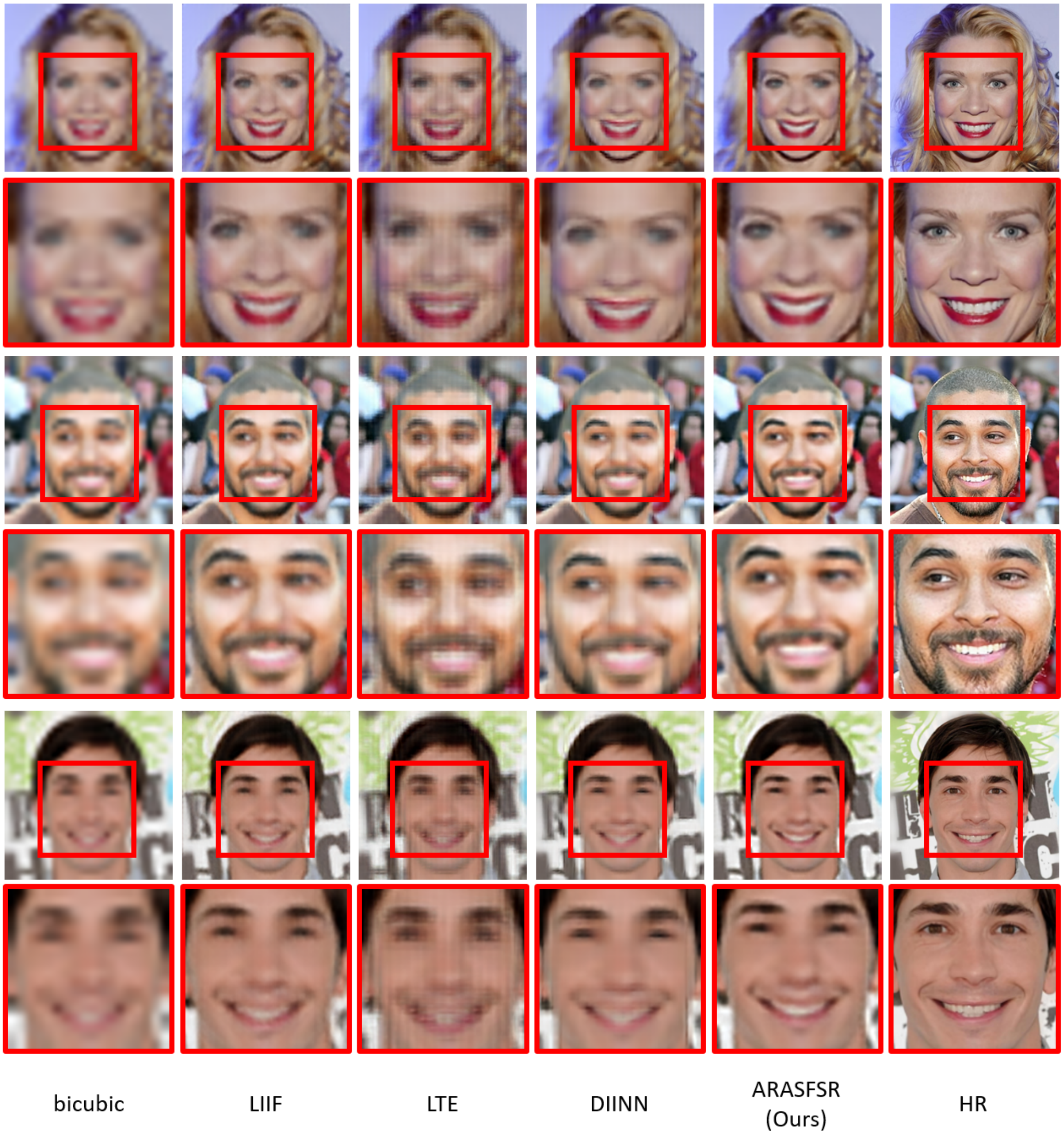

Figure 2. Visual comparison on CelebAHQ with INR-based SISR methods (LR-HR: 32-256).

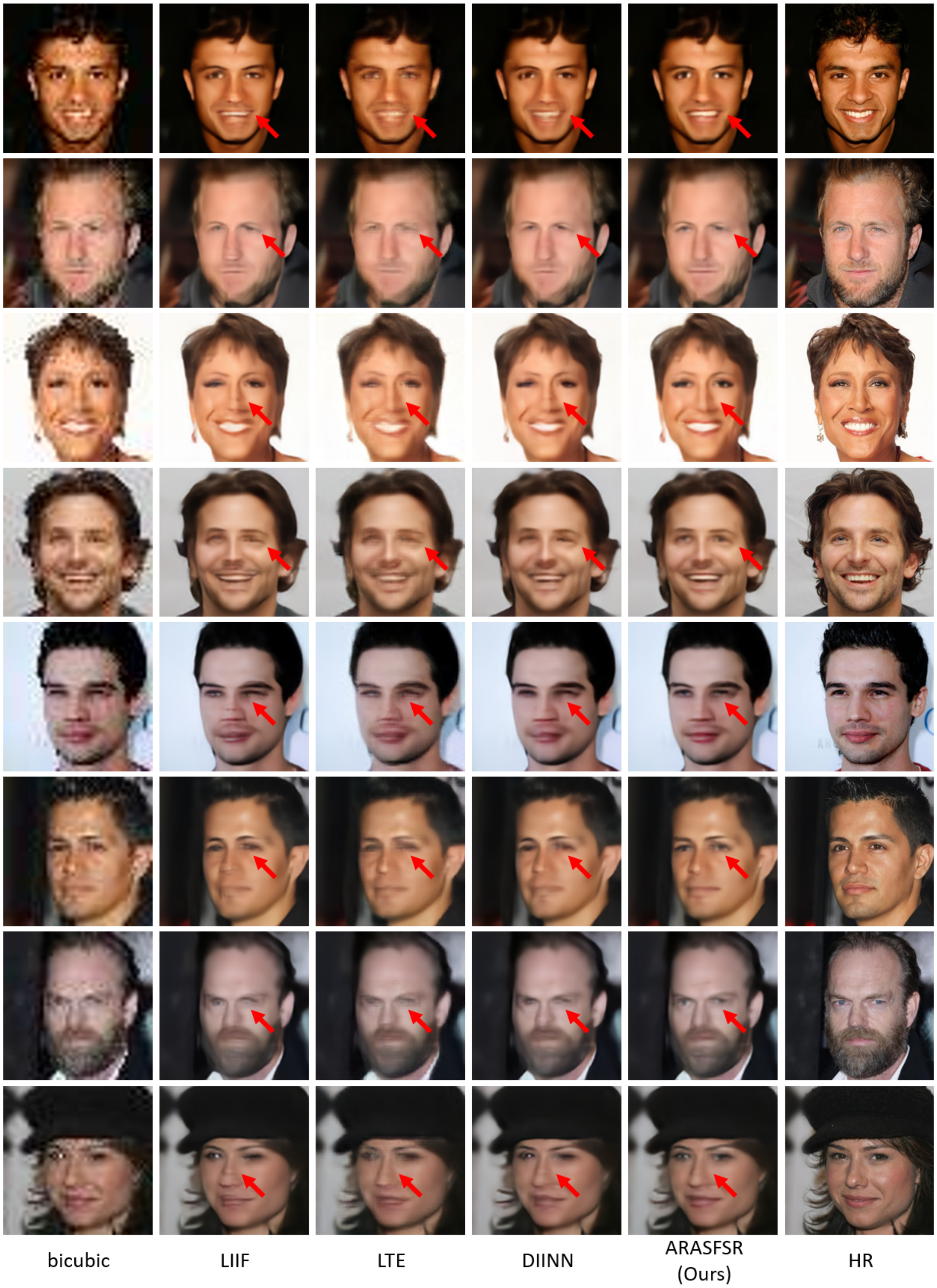

Figure 3. Visual comparison of real-world case on CelebAHQ-NN-JPEG with INR-based SISR methods (LR-HR: 48-256).

receptive-field features. In *International Conference on Artificial Neural Networks,(ICANN2022)*, 2022. 1

[3] 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. 2

[4] 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. 1

[5] 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. 1

[6] 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. 2

[7] 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. 1