Multi-Reference Image Super-Resolution: A Posterior Fusion Approach

Ke Zhao, Haining Tan, and Tsz Fung Yau

Abstract—Reference-based Super-resolution (RefSR) approaches have recently been proposed to overcome the ill-posed problem of image super-resolution by providing additional information from a high-resolution image. Multi-reference super-resolution extends this approach by allowing more information to be incorporated. This paper proposes a 2-step-weighting posterior fusion approach to combine the outputs of RefSR models with multiple references. Extensive experiments on the CUFED5 dataset demonstrate that the proposed methods can be applied to various state-of-the-art RefSR models to get a consistent improvement in image quality.

Index Terms—Reference-based Super-Resolution, Image Fusion, Adaptive Weight Masking

✦

1 INTRODUCTION

Single-image super-resolution (SISR) is a computer vision task that reconstructs a high-resolution (HR) image from a low-resolution (LR) image. Typically, estimating a high-resolution image from its low-resolution counterpart is an ill-posed inverse problem [1], meaning that there are infinitely many solutions that satisfy the measurements. This underdetermined nature of the problem is particularly pronounced for images with abundant high-frequency details. To reach an optimal solution to the inverse problem with respect to certain criteria, additional regularization terms need to be specified. However, no simple regularization term can be specified to cover the characteristics of all kinds of images, thus conventional SISR algorithms are usually poorly performed.

Reference-based super-resolution (RefSR) methods explicitly exploit additional information from an external HR reference image to enhance the SISR process. Intuitively, sufficient information is encoded in a reference image that contains the same content as that on the LR image to facilitate texture restoration. However, a majority of current RefSR models can only take one reference image, limiting the amount of supplementary information to incorporate into the super-resolution process.

To achieve multi-reference-based super-resolution (MRefSR), two approaches are possible: it can either be that multiple reference images are used as the initial inputs to the model, or that the multiple outputs of SRefSR using different reference images are fused to combine the information. We observed that nowadays most RefSR models aim to achieve better content alignment, both spatially and semantically, between the input image and one reference image. Because it is intuitively hard for additional reference images to contribute to the alignment process, we claim that the posterior fusion of multiple SRefSR outputs would be a more natural way to combine the relevant information from each reference image. Following this idea, we proposed a two-step-weighting fusion scheme that can be incorporated into a variety of existing SRefSR models to achieve MRefSR and better-quality final SR images. Also, our proposed method has a low computational cost and allows for parallel computation for the super-resolution process with multiple reference images.

The proposed posterior fusion method can be applied to a wide range of applications. For example, in video game graphic rendering, HR patches for each object in the scene are readily available as the texture to be mapped, and they can be used as reference images to perform MRefSR. This approach would be applied to any video game and would save huge computing resources compared to the state-of-the-art NVIDIA DLSS, which trains separate SR neural networks for each video game.

2 RELATED WORK

2.1 Single Image Super-Resolution

With the popularity of Convolutional Neural Networks (CNN), learning-based approaches demonstrate significantly better performances given an appropriate training set. Early-stage CNN-based SISR models like SRCNN [2] choose pixel-level reconstruction errors such as MSE and MAE between the recovered HR image and ground truth as loss functions to optimize. Furthermore, significant improvements can be made by optimizing the standard CNN architecture. For instance, the approach EDSR [3] proposed by Lim et al. applied the residual network architecture to the SR task and achieved superior results. While these algorithms tend to maximize the peak signal-to-noise ratio (PSNR), they often result in smooth reconstruction lacking high-frequency details and are perceptually unsatisfying. To state the problem of SISR in another way, downsampling an HR image is an irreversible compression process during which much high-frequency information is lost. Instead of trying to recover the lost information from nowhere, Ledig et al. [4] adopt Generative Adversarial Networks (GAN) and proposed SRGAN that generates "fake" texture details that are visually realistic. While these results are perceptually satisfying, texture details in these images are hallucinations and are often different from those in the ground-truth images, resulting in PSNR degradation. This deficiency makes the methods like SRGAN unsuited for fidelity-sensitive applications like medical imaging. Additionally, pure GAN-based SISR approaches fail to produce satisfying results on

test images with complicated components, often resulting in distorted color lumps.

2.2 Single-Reference-based Super-Resolution

Zheng et al. [5] proposed an end-to-end approach, named CrossNet, based on fully convolutional neural networks that can perform spatial alignment between the reference features and the LR features. One issue of this model is that regions of the reference image that are irrelevant to the input will degrade the performance. Motivated by this, Shim et al. [6] proposed a robust RefSR model that is aware of the relevancy of the reference image, leading to a more robust result that outperforms SRGAN in terms of generating visually stratifying SR images while also achieving high PSNR. More recently, Zhang et al. [7] proposed to use dual zoomed observations (from a telephoto) as references and apply self-supervised techniques to that. This is inspired by multiple cameras in modern smartphones that are able to collect dual-zoomed observations at the same time. While this model performs well for scenes with repeated structural texture, it gives highly distorted outputs for common scenarios.

2.3 Multi-Reference-based Super-Resolution

MRefSR has recently been proposed to extend the idea of RefSR, and previous works in this field majorly focus on designing novel neural network models such that they take multiple reference images as the initial inputs. Yan et al. [8] proposed a content-independent MRefSR model that builds up a universal reference pool before doing predictions. Given an input low-resolution image alone, this model finds reference images whose textures are similar to each segmentation of the input from its pool to help the reconstruction process. A hierarchical attention-based sampling approach is proposed by Pesavento et al. [9] to combine the features of multiple reference images. While these models provide a promising direction, the performance improvement compared to SRefSR is still limited.

3 PROPOSED METHOD

The overview of our proposed multi-image RefSR pipeline is shown in Fig. 1. This proposed method consists of two major parts: 1) An image alignment and texture extraction module, which can be any existing single-image RefSR model, and 2) A image fusion module. The first module takes an LR input image and an HR reference image and tries to match the HR reference spatially and semantically to the LR input and then uses the corresponding HR texture in the reference image to get an HR version of the input image. By feeding one input LR image and multiple reference images to this module, multiple SR output images are obtained. The second module fuses these output images into a single SR image by combining the best region of each SR output, with the objective to improve image quality. The fusion module consists of two steps, namely adaptive weight masking and globally reference-quality-based weighted averaging.

3.1 Naive Fusion

Before introducing our proposed fusion method, let's first consider the simplest fusion scheme, which is just averaging the intensity of pixels of each single-reference SR output. That is, for every pixel position p in the SR image $\mathbf{I}_i$ generated by the RefSR module, the corresponding pixel value of the fused image $\hat{\mathbf{I}}$ is given by:

$$\hat{\mathbf{I}}(p) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_i(p) \tag{1}$$

where N is the total number of reference images and $\mathbf{I}(p)$ is the intensity of the pixel at position p. The index i of SR image $\mathbf{I}_i$ is arranged such that the lowest indexed image $\mathbf{I}_1$ corresponds to the RefSR result with the most relevant reference image.

By relevant we mean the content of the reference image is similar to that of the input image. Take the CUFED5 dataset that is adopted for evaluation by this paper as an example, the leftmost image in Fig. 3 is the ground truth HR version of the LR input image, while the other images are the reference images whose relevance to $\mathbf{I}_{GT}$ decreases from left to right.

It is intuitive that $\mathbf{I}_1$ would have the best quality and $\mathbf{I}_N$ would have the worst. As most of the previous single-reference SR works take the $\mathbf{I}_1$ as their final results, averaging $\mathbf{I}_1$ with the rest $\mathbf{I}_i$ will degrade the quality of the final fused image $\hat{\mathbf{I}}$. These claims are supported by our experimental results.

3.2 Adaptive Weight Masking

Instead of naively averaging the single-reference SR image $\mathbf{I}_i$ as a whole, a more desirable fusion method would combine the best of each SR image. As illustrated in Fig. 1, even though $\mathbf{I}_1$ (the top one in the middle column) has the best overall quality, the $\mathbf{I}_3$ has sharper character reconstruction on the selected region. To achieve this, an adaptive weight mask $\mathbf{W}_i$, whose dimension is the same as SR image $\mathbf{I}_i$, is computed. It gives higher weights to pixels in regions with better-quality reconstruction and each pixel value of the fused image $\hat{\mathbf{I}}$ is given by:

$$\hat{\mathbf{I}}(p) = \frac{1}{\sum_{i=1}^{N} \mathbf{W}_i(p)} \sum_{i=1}^{N} \mathbf{I}_i(p) \mathbf{W}_i(p) \tag{2}$$

Ideally, $\mathbf{W}_i(p)$ should measure how close a pixel is to the ground-truth image, but this information is not available in real RefSR scenarios. What is readily available is the LR input image $\mathbf{I}_{input}$ itself.

To compute $\mathbf{W}_i(p)$, $\mathbf{I}_i$ is first downsampled to have the same dimension as the input image, denoted by $\mathcal{D}(\mathbf{I}_i)$. The idea is that, for an ideal reconstruction of the input image, the downsampled version of it would be exactly the same as the input image. Therefore the difference between the pixel intensities of $\mathcal{D}(\mathbf{I}_i)$ and $\mathbf{I}_{input}$ would be a good proximity to the difference between $\mathbf{I}_i$ and $\mathbf{I}_{GT}$. Specifically, $\mathbf{W}_i$ is given by:

$$\mathbf{W}_i = \mathcal{U}(\exp\left(-\beta(\mathcal{D}(\mathbf{I}_i) - \mathbf{I}_{input})^2\right)) \tag{3}$$

where $\mathcal{U}$ denotes bicubic upsampling and exp is element-wise exponential. β is a parameter controlling how much the discrepancy in pixel intensities is penalized.
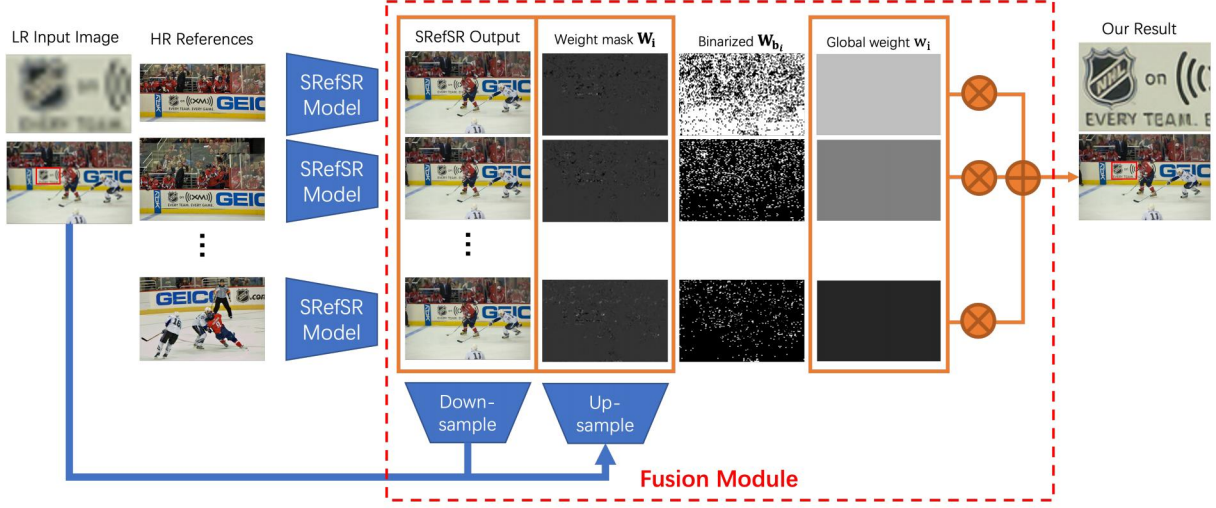
Fig. 1. Our proposed multi-image RefSR pipeline

3.3 Global Reference-Quality-based Weight

While the Adaptive Weight Masking scheme works reasonably well in cases for severely distorted regions in $\mathbf{I}_i$, it is insensitive to small distortions in $\mathbf{I}_i$, which is crucial in fine detail recovery. To see why this is happening, consider the most naive super-resolution result of $\mathbf{I}_{iput}$, that is, its bicubic interpolated upsampled version. Despite having the worst SR quality, this image would be yet another optimal solution with maximized $\mathbf{W}_i$ since its bicubic downsampled version would be exactly the same as the $\mathbf{I}_{iput}$.

To overcome this problem, an measurement to encourage fidelity is needed. One might consider Natural Image Prior [10] as a way to force fine details. However, we have observed the deficiencies in the RefSR could be both blurring lumps and noise high-frequency mosaic, making Natural Image Prior noneffective. Instead, we took an indirect measurement of the fidelity of $\mathbf{I}_i$. As shown in Fig. 2, the black-and-white figure is the binary weight mask computed by finding the maximum value of $\mathbf{W}_i(p)$ for each pixel across all RefSR results $\mathbf{I}_i$, so the pixel intensity each the binary weight mask is given by:

$$\mathbf{W}_{b_i}(p) = \begin{cases} 1 & , i = \underset{i}{\operatorname{argmax}} \, \mathbf{W}_i(p) \\ 0 & , \text{otherwise} \end{cases} \quad (4)$$

It can be observed from Fig. 2 that the total area of the white region in the binary weight mask figure can be an indicator of how relevant the underlying reference image is to the input image. Intuitively, the better the underlying reference, the better the texture quality in the SR results. Therefore, the sum of the binary weight mask is adopted as a global weight for $\mathbf{I}_i$ in the fusion process. This global weight is computed by:

$$w_i = \exp\left(\beta_g \sum_p \mathbf{W}_{b_i}(p)\right) \quad (5)$$

And the fused image is given by:

$$\mathbf{I}_{fused} = \frac{1}{\sum_i w_i} \sum_i w_i \hat{\mathbf{I}}_i \quad (6)$$

where $\hat{\mathbf{I}}_i$ is the ith SR image after applied Adaptive Weight Masking, and β_g is a parameter controlling how much priority is given to the $\hat{\mathbf{I}}_i$ with the best underlying referencing image.

Fig. 2. Binary weight mask computed from RefSR results

4 EXPERIMENTAL RESULTS
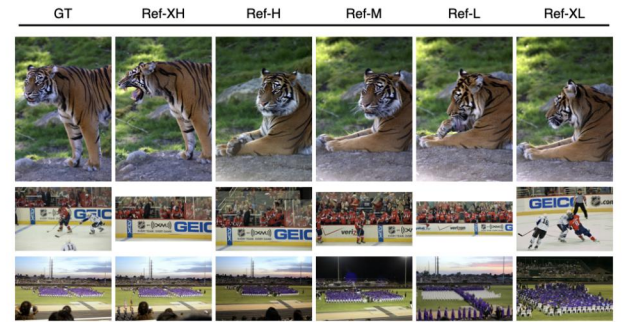
4.1 Dataset and Evaluation Metrics

Fig. 3. Sample images from the test set of CUFED5

Testing Dataset. The test set of CUFED5 [11] is used as the testing dataset because it provides multiple reference images for each low-resolution image. (Fig. 3). The CUFED5 test set has 126 HR input images and each has 5 HR reference images with different similarity levels. The input LR images for the evaluation below are constructed by 4x bicubic downsampling from the ground-truth HR images.

Evaluation Metrics. The quantitative experiments adopt PSNR and SSIM (structural similarity) [12] on the Y channel

of the YCrCb space as evaluation metrics. In case that different RefSR modules generate images with different dimensions, the cropping/padding/interpolation scheme in the fusion module is chosen to be the same as that in the RefSR modules so that the evaluation results are consistent.

4.2 Qualitative Comparisons

Table 1 summarizes the qualitative comparison with the state-of-the-arts. We applied our method on C^2-Matching [13] and AMSA [14] and compared the results with the original ones. Also the result of ESRGAN [15] is included as a representation of SISR results. It can be observed that our method does a certain level of visual quality improvement upon the original work, especially in the case of C^2-Matching, where a denoising bonus is applied upon super-resolution.

4.3 Quantitative Evaluation

Overall Comparisons. Table 4.3 shows the quantitative comparison of the performance of a variety of super-resolution methods. We applied the proposed method to C^2-Matching [13] and AMSA [14], and compared and results with the original works. We also include the results of representative models of SISR and RefSR, which are evaluated on the same dataset as outs. For SISR methods, we include SRCNN [2], EDSR [3], RCAN [16], SRGAN [4], ENet [17], ESRGAN [15] and RankSRGAN [18]. For RefSR methods, we include CrossNet [5], SRNTT [11], TTSR [19], SSEN [13], E2ENT^2 [20] and CIMR [8].

We can see that our method outperforms all SISR and Ref SR models. In particular, C^2-Matching and AMSA have shown improvement after integrating with the pipeline.
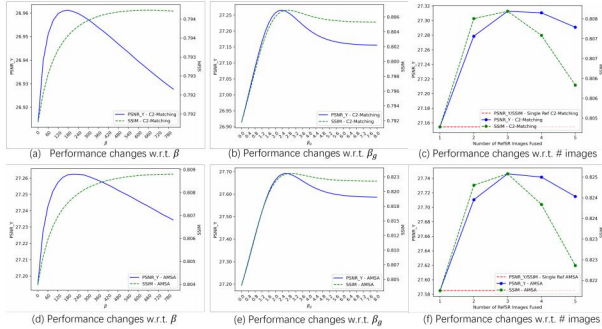

Fig. 4. Quantitative evaluation by applying the proposed method to C^2-Matching and AMSA

Evaluate Adaptive Weight Masking. We applied the proposed method to C^2-Matching and AMSA to perform quantitative evaluations. To analyze the effectiveness of Adaptive Weight Masking, we fixed β_g to be 0, which essentially disables the effect of the Global Reference-Quality-based weight. Fig. 4 (a) and (d) shows how the PSNR_Y/SSIM changes as β varies from 0 to 810. Note that when $\beta = 0$, the method degrades to Naive Fusion. It can be observed that both PSNR_Y and SSIM get better as β increase, that is, as the distorted pixels are penalized more and more. While PSNR_Y begins to decrease as β passes the optimal value, SSIM keeps increasing and then decreases mildly compared with PSNR_Y. While the increased PSNR_Y/SSIM
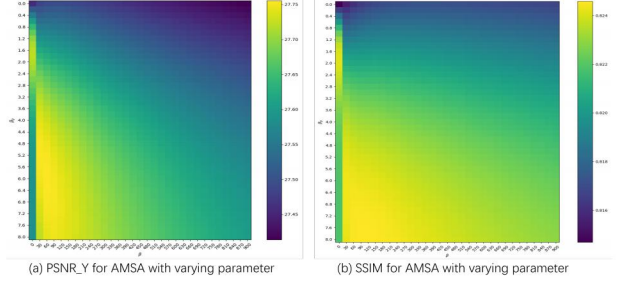

(a) PSNR_Y for AMSA with varying parameter (b) SSIM for AMSA with varying parameter

Fig. 5. The PSNR_Y and SSIM heatmap with varied parameters β and β_g

validates the effectiveness of Adaptive Weight Masking, it should be noticed that the PSNR_Y/SSIM is worse than the 27.16/0.805 (see Table 4.3) achieved by C^2-Matching and xx achieved by AMSA using the single most relevant reference image. That is due to the Adaptive Weight Masking's insensitiveness to relatively small distortions in SR results, as stated in Sec. 3.3.

Evaluate Global Reference-Quality-based Weight. As shown in Fig. 4 (b) and (e), as β_g increases from 0 to 8 the PSNR_Y/SSIM of both C^2-Matching and AMSA first arise then plateau, given the condition that β is fixed to 0. When $\beta_g = 0$ the process is equivalent to Naive Fusion while $\beta_g \to \infty$ would be the case of using the single best SRefSR result alone. This trend shows that there is indeed additional valuable information from the sub-optimal SRefSR results, and even a simple trick as weighted averaging would improve the image quality with a properly chosen β_g.

Evaluate the Combined Weighting Method Fig. 5 how the metrics values changes with varied β and β_g, and only the results for ASMA are demonstrated since the evaluation with C^2-Matching produce similar results. It can be observed there is a sharp change when β changes from 0 to 30, which is consistent with the previous observation on the quick increase in metrics values in Fig. 4 (a) and (d).

Evaluate the Number of Images Fused. In Fig. 4 (c) and (f), we cumulatively fuse more and more RefSR images to and show the resulting image's PSNR_Y/SSIM. Notice that the order of fusion is chosen such that the best-quality RefSR image is taken to be fused with RefSR images that are of less and less quality. It can be shown that our proposed method is resistant to imperfect RefSR images since the PSNR_Y and SSIM starts to decrease only after the fourth RefSR image is fused and PSNR_Y decrease mildly.

It is noteworthy that the evaluation above is done separately with two different state-of-the-art SRefSR models, namely C^2-Matching and AMSA, and a consistent behavior is observed. Therefore we claim that the proposed method has the potential to be incorporated into a variety of SRefSR pipelines to get a performance improvement.

4.4 Case Study with SelfDLSR

The models in the previous experiment use HR references to generate super-resolution images, so we have also studied the effectiveness of this super-resolution pipeline with other types of references. In particular, we experimented with Self-DZSR which uses telephotos (zoom-in images) as references.
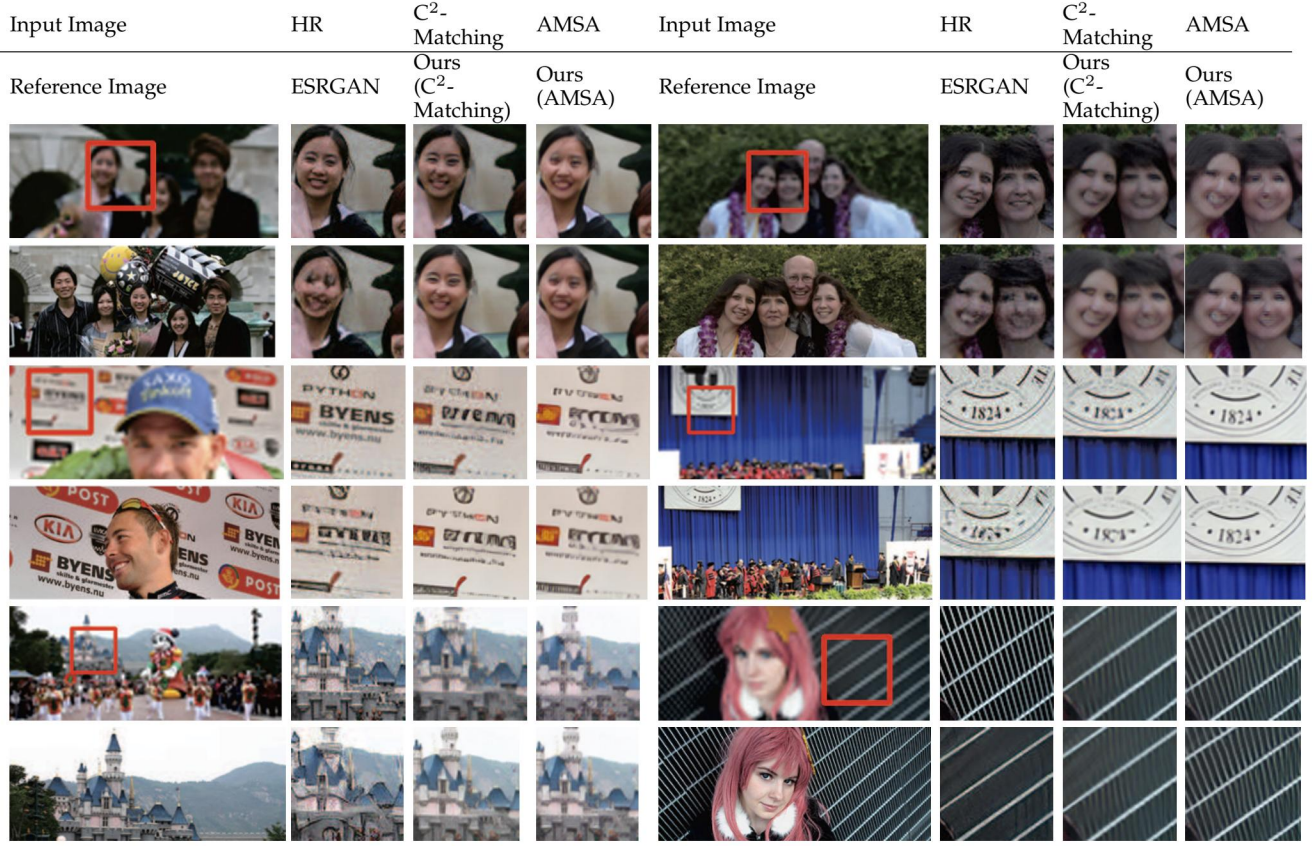
Input Image	HR	C^2-Matching	AMSA	Input Image	HR	C^2-Matching	AMSA
Reference Image	ESRGAN	Ours (C^2-Matching)	Ours (AMSA)	Reference Image	ESRGAN	Ours (C^2-Matching)	Ours (AMSA)

TABLE 1
Qualitative results

TABLE 2
Evaluation of the models on PSNR_Y and SSIM

	Model	PSNR/SSIM
SISR	SRCNN	25.33/0.745
	EDSR	25.93/0.777
	RCAN	26.06/0.769
	SRGAN	24.40/0.702
	ENet	24.24/0.695
	ESRGAN	21.90/0.633
	RankSRGAN	22.31/0.635
Ref SR	CrossNet [5]	25.48/0.764
	SRNTT	25.61/0.764
	TTSR	25.53/0.765
	SSEN	25.35/0.742
	E2ENT^2	24.01/0.705
	CIMR	26.16/0.781
	C^2-Matching	27.16/0.805
	AMSA	27.31/0.809
	Ours (with C^2-Matching)	**27.29/0.806**
	Ours (with AMSA)	**27.56/0.825**

Dataset Preparation. While the CUFED5 test set can be directly adopted for AMSR and C^2-matching, it does not contain telephotos that are required for the DZSR model. Therefore, an image processing pipeline is constructed to generate short-focus and telephotos from CUFED5. This takes the ground truth high-resolution images alone and outputs the short-focus low-resolution images by 4x bicubic downsampling, the same as the previous experiment with C^2-Matching and AMSA. It also renders telephotos by cropping the high-resolution images to simulate the effect of zooming in. This pipeline is flexible because it can support any image dataset and different resize factors for more comprehensive comparisons.

Fig. 6. Qualitative Comparisons by using different reference images (telephoto) for SelfDZSR. The leftmost one is the fused image while the right 5 are outputs from individual references.

Qualitative Comparisons. Figure 6 shows the qualitative results. When center reference (telephoto) are used, we can see that the resulting images show obvious artifacts in the surrounding of the output images. However, when we use the telephotos taken from the corners of short-focused images, the center part of the super-resolution images seems to be replaced by the reference, largely deviating from the ground truth. This is different from what we expected as the reference patches used during training are not restricted to

the center but at randomized positions with simple augmentation (flipping and $90°$ rotation), making it resilient to the displacement of reference. Note that the surrounding parts are smoother without the artifacts. Nonetheless, the fused output (the leftmost one) can combine the smoothness from non-center references and the overall structure with center references, showing significant improvement from each of the single-reference super-resolution outputs.

Fig. 7. Binary weight masks from Global Reference-Quality-based Weight by using different reference images (telephoto) for SelfDZSR.

Weight Analysis. Figure 7 illustrates the binary weight masks for different references respectively, we can see that in the first reference (using a center telephoto), most of the pixels in the center part have the highest weights across references, showing that the Adaptive Weight Masking step in the pipeline can identify the distortion in center parts and put lower weights. We also discover that some of the pixels in the surrounding area are the highest in non-center reference (the second to fifth in the figure), showing that the step can recognize the artifacts in the surroundings. Therefore, the fusion module can indeed combine the smoothness from non-center references and the overall structure with center references.

5 DISCUSSION

To improve our image fusion module, we will try to explore better ways to assign pixel and image weights for Adaptive Weight Masking and Global Reference-Quality-base. In particular, we will adopt other priors that combine the sharp region of each image to provide a more fine-grained result. This can be gradient-based methods such as Poisson image processing and pyramid-based methods such as Gaussian pyramid.

We can also adopt neural network models for image fusion such that the prior knowledge of what a "natural image" is learned by the neural network to select the best-reconstructed regions in each RefSR image. This approach has the potential to significantly outperform our simple global reference-quality-based weighting strategy.

REFERENCES

[1] S. Kathiravan and J. Kanakaraj, "An overview of sr techniques applied to images, videos and magnetic resonance images," *Smart CR*, vol. 4, no. 3, pp. 181–201, 2014.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *CoRR*, vol. abs/1501.00092, 2015. [Online]. Available: http://arxiv.org/abs/1501.00092

[3] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[4] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[5] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[6] G. Shim, J. Park, and I. S. Kweon, "Robust reference-based super-resolution with similarity-aware deformable convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Z. Zhang, R. Wang, H. Zhang, Y. Chen, and W. Zuo, "Self-supervised learning for real-world super-resolution from dual zoomed observations," *arXiv preprint arXiv:2203.01325*, 2022.

[8] X. Yan, W. Zhao, K. Yuan, R. Zhang, Z. Li, and S. Cui, "Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation," in *European conference on computer vision*. Springer, 2020, pp. 52–68.

[9] M. Pesavento, M. Volino, and A. Hilton, "Attention-based multi-reference learning for image super-resolution," Aug 2021. [Online]. Available: https://arxiv.org/abs/2108.13697

[10] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[11] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Y. Jiang, K. C. Chan, X. Wang, C. C. Loy, and Z. Liu, "Robust reference-based super-resolution via c2-matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2103–2112.

[14] B. Xia, Y. Tian, Y. Hang, W. Yang, Q. Liao, and J. Zhou, "Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution," *arXiv preprint arXiv:2201.04358*, 2022.

[15] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[16] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[17] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[18] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[19] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5791–5800.

[20] Y. Xie, J. Xiao, M. Sun, C. Yao, and K. Huang, "Feature representation matters: End-to-end learning for reference-based image super-resolution," in *European Conference on Computer Vision*. Springer, 2020, pp. 230–245.