This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dual-Camera Super-Resolution with Aligned Attention Modules

Tengfei Wang[1*] Jiaxin Xie[1*] Wenxiu Sun[2] Qiong Yan[2] Qifeng Chen[1]
[1]HKUST [2]SenseTime Research and Tetras.AI

Abstract

We present a novel approach to reference-based super-resolution (RefSR) with the focus on dual-camera super-resolution (DCSR), which utilizes reference images for high-quality and high-fidelity results. Our proposed method generalizes the standard patch-based feature matching with spatial alignment operations. We further explore the dual-camera super-resolution that is one promising application of RefSR, and build a dataset that consists of 146 image pairs from the main and telephoto cameras in a smartphone. To bridge the domain gaps between real-world images and the training images, we propose a self-supervised domain adaptation strategy for real-world images. Extensive experiments on our dataset and a public benchmark demonstrate clear improvement achieved by our method over state of the art in both quantitative evaluation and visual comparisons. Our code and data are available at https://tengfei-wang.github.io/Dual-Camera-SR/index.html.

1. Introduction

Most smartphone manufacturers adopt an asymmetric-cameras system consisting of multiple fixed-focal lenses instead of a variable-focal one for optical zoom, due to limited assembly space. As shown in Fig. 1, the most common configuration has dual cameras with wide-angle (main camera) and telephoto lenses that have different field of views (FoV). The wide-angle and telephoto images often have spatial misalignment and color discrepancy due to viewpoint differences and different image signal processing (ISP) pipelines in the two lenses. As these two images capture the same scene with different focal lengths, can we use the telephoto image as a reference to enhance the resolution of the wide-angle image? To answer this question, we study reference-based super-resolution (RefSR) with the focus on dual-camera super-resolution (DCSR).

The key challenges of RefSR lie in (1) *how to effectively establish correspondences between low-resolution inputs (LR) and reference images (Ref)* (**feature warping**), and (2) *how to integrate the reference information to improve the output image quality* (**feature fusion**). It has been widely observed that similar semantic patches and texture patterns tend to recur in the same or highly-correlated images with variable positions, orientations and sizes [23, 46]. To search and utilize these correlated patterns from reference images, previous learning-based approaches adopt either patch-wise matching (patch-match [44, 43], patch-based attention [34, 35]) or pixel-wise alignment (optical-flow [45], offsets [28]), with different pros and cons. The pixel-wise alignment is able to handle non-rigid transformation, but usually less stable and prone to generate distorted structures due to the difficulty of reliable flow or offsets estimations [6], especially for largely misaligned reference images. Patch-wise matching can achieve compelling warping performance since it evaluates similarity scores between LR and Ref patches in an explicit fashion. However, the vanilla patch-level matching lacks robustness to spatial misalignment, e.g. scaled or rotated patches. As shown in Fig. 1, even though highly-similar patches are available in the reference image, previous approaches are insufficient to make use of these cues, and tend to average the misaligned Ref and LR patches to produce blurry images.

Another limitation of previous RefSR approaches is that they are difficult to be directly applied to high-resolution images captured by smartphones. The reference images in RefSR datasets [43] are typically smaller than 512×512. Most methods thus globally searches over the entire reference image for super-resolution cues. Nevertheless, the memory consumption of a global searching strategy would be intractable for the high-resolution cases (e.g. 4K). The domain gaps between real-world images and training images can also degrade the zoom performance [13, 40, 5].

To tackle these issues, we propose a deep RefSR method with the focus on dual-camera super-resolution. First, we generalize the vanilla patch-based attention to an aligned attention module, which searches for related patches based on explicit matching, while implicitly learning inter-patch transformations to alleviate spatial misalignment. Second, to prevent the reference patches from idling and contributing less to the super-resolution results, we impose a fidelity

[1]equal contribution

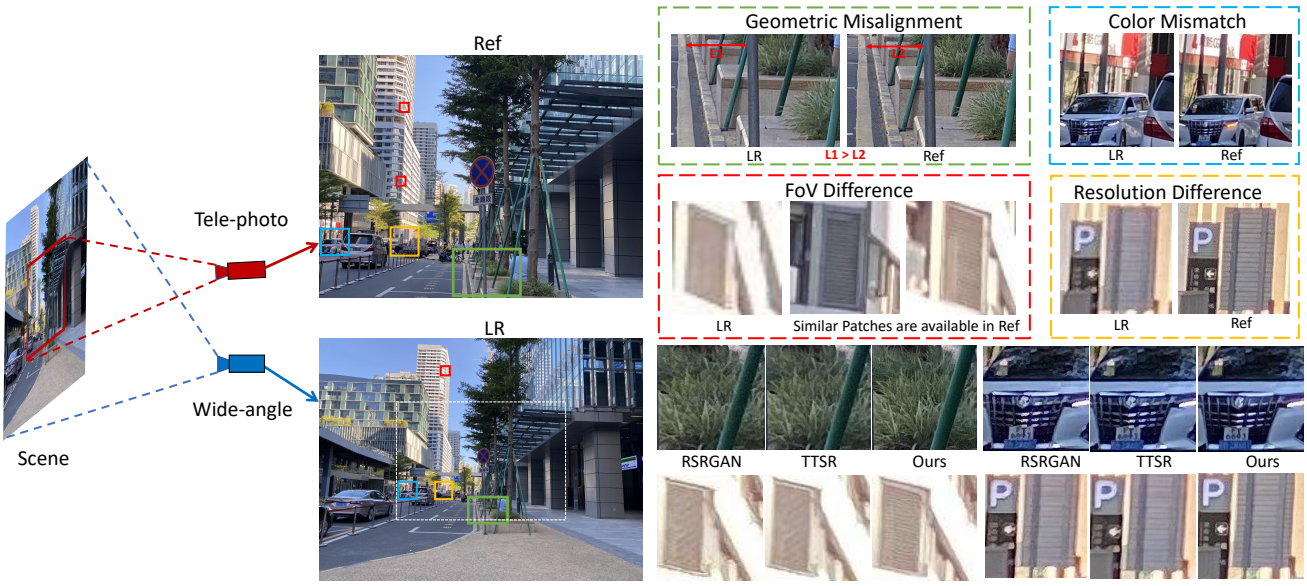

Figure 1. Demonstration of the smartphone dual-camera system. The telephoto and wide-angle images share similar contents within the overlapped FoV (indicated by the white dotted line), while various misalignment exists. We take the telephoto image as the reference to super-resolve the wide-angle image for combining both large FoV and high-quality details. Compared with state-of-the-art SR approaches RSRGAN [39] and TTSR [35], our results are sharper and more realistic. Zoom-in for details.

loss on the reference images. To advance our method to real-world images, we also propose a self-supervised adaptation strategy. The main contributions of our paper can be summarized as:

- We are the first to explore the real-world dual-camera super-resolution (wide-angle and telephoto cameras). We propose a self-supervised domain adaptation scheme to bridge domain gaps between real-world images and downsampled images.
- We propose the aligned attention module and adaptive fusion module to improve the RefSR architecture. Our method outperforms state-of-the-art approaches qualitatively and quantitatively.
- We argue the importance of imposing an explicit fidelity loss on reference images and performing explicit high-frequency fusion in the image space to the super-resolution quality.

2. Related Work

2.1. Single Image Super Resolution

SISR [11] has been actively explored in recent years. After SRCNN [9], MDSR [18] introduced residual blocks to super-resolution area. RCAN [41] further improved residual blocks by channel attention. To improve the perceptual quality, Johnson et al. [15] proposed the perceptual loss to minimize the feature distance. SRGAN [17] adopted generative adversarial networks [12] for more realistic textures. ESRGAN [32] enhanced SRGAN with Residual-in-Residual Dense Block. RankSRGAN [39] combined SRGAN with a well-trained ranker that gives ranking scores. CSNLN [23] proposed cross-scale non-local attention to find self-similarity for high-quality reconstruction.

2.2. Reference-based Super Resolution

RefSR alleviates the ill-posed nature of SISR by providing high-resolution reference images. Previous learning-based approaches adopt either patch-wise matching or pixel-wise alignment for feature warping. Pixel-wise alignment methods usually build a dense corresponding map, and warp the reference feature maps pixel by pixel. Zheng et al. [45] proposed to estimate optical flow between input and reference images to warp feature maps at different scales. However, it remains a challenging problem for reliable flow estimation in largely misaligned regions. Shim el al. [28] proposed to implicitly estimate the offsets with deformable convolution[8] instead of optical flow. The offsets warping is faster and more flexible than the flow counterpart, while it is typically less stable.

Patch-wise matching searches for related patches by calculating similarity scores explicitly, which is thus more stable with better interpretability. Zhang et al. [43] adopted Patch Match [2] to warp features extracted by a pretrained VGG network [29]. With a fixed VGG network as feature extractors, their method does not train the extractor jointly with the reconstruction net. Yang et al. [35] and Xie et al. [34] further proposed to adopt a learnable extractor and replace Patch Match with a patch-based attention, which allows an end-to-end learning pipeline. These patch-level warping methods can find semantic-similar patches, but are

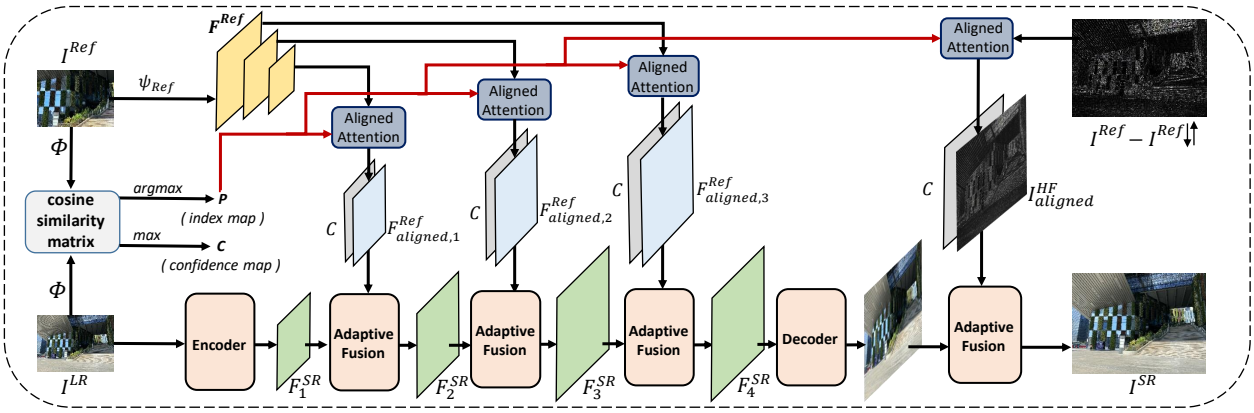

Figure 2. Overview of our approach. We first match the nearest I^{Ref} patch for each I^{LR} patch under cosine distance in feature space Φ. The reference features F^{Ref} at different scales are then warped by the aligned attention and adaptively fused with SR feature F^{SR} according to this matching. After the final fusion with high-frequency $I^{HF}_{aligned}$, the network yields a high-fidelity output I^{SR}.

non-robust to inter-patch misalignment (e.g. scaled and rotated patches), which typically leads to blurry results. To address this issue, we proposed the aligned attention module that robustly warps spatially misaligned patches by estimating patch-wise alignment parameters.

2.3. Dual Camera Super Resolution

The dual-camera super-resolution aims at super-resolving the wide-angle image with the telephoto image as a reference, which combines both large FoV of short-focal camera and high resolution of long-focal camera. Most related works adopt traditional global correctness and registration techniques. Park et al. [25] and Liu et al. [19] assumed that there is no disparity between wide-angle and telephoto pairs, and only correct brightness and color globally. Some prior work considered the geometric misalignment between inputs and references. They simulated tele-image by center-cropping HR and performing random affine transformation, and formulate this task to image registration. Yu et al. [37] applied RANSAC algorithm [10] on SURF [4] features to conduct global registration. Manne et al. [20] applied FLANN algorithm [24] on ORB features [26] for geometric registration. Nevertheless, there are huge domain gaps between the real-world telephotos and the simulated ones [13, 5], and previous approaches usually show significant performance drop in the practical configuration. Instead of global image registration, We formulate DCSR as a setting of RefSR, and propose an end-to-end pipeline and training strategy. To the best knowledge of ours, we are the first learning-based method for real-world dual-camera super-resolution.

3. Method

Given I^{LR} and I^{Ref}, we aim at generating a high-resolution image I^{SR} that possesses high-quality details conditioned on I^{Ref}. As shown in Fig. 2, our end-to-end pipeline consists of two parts: feature warping with aligned attention modules (Section 3.1), and feature fusion with adaptive fusion modules (Section 3.2). To utilize both high-level and low-level information provided by I^{Ref}, following previous works [43, 35], we extract reference features F^{Ref} at different scale levels via encoder ψ_{Ref}. At each scale, we perform the aligned attention on F^{Ref} to warp it to match the LR for later fusion. This module can robustly match correlated patches and further align these patches to alleviate the differences in orientations and scales. After that, the aligned features $F^{Ref}_{aligned}$ as well as the high-frequency residual $I^{HF}_{aligned}$ are sequentially integrated with LR information guided by the matching confidence.

3.1. Feature Warping with Aligned Attention

Our method stems from the observation that similar patches tend to recur across correlated images with different scales and orientations [3]. The aligned attention aims at searching for these related reference patches and warp them to align with the LR counterparts. Following [43], we first perform a patch-wise matching [7] to coarsely warp the reference, which is briefly reviewed below. $I^{LR}\uparrow$ and I^{Ref} are first embedded into feature maps via a shared encoder $\phi(\cdot)$, and densely (stride=1) divided to 3×3 patches, where $\uparrow$ denotes bicubic upsampling. We then calculate the cosine distance $S_{i,j}$ between each pairs of LR-patch i and Ref-patch j. For each LR-patch, we want to select the most relevant Ref-patch for later feature fusion. The index map P and confidence map C of the matching are obtained as:

$$P_i = \arg\max_j S_{i,j}, \qquad C_i = \max_j S_{i,j}. \tag{1}$$

The index map indicates the most relevant Ref-patch-P_i for each LR-patch-i, and the confidence map gives the

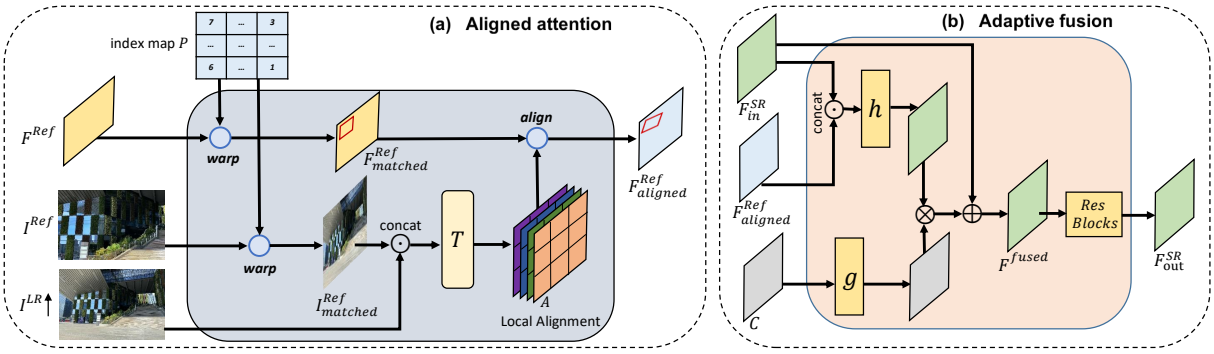

Figure 3. Illustration of the aligned attention module and adaptive fusion module. (a). The aligned attention applies index map P to coarsely warp F^{ref} and then fine-aligns the patch with the learned local transformation. (b). The adaptive fusion applies an additional convolution layer g to aggregate neighbor information in the confidence map C.

matching confidence C_i for this match. The reference patches can be warped now according to the index map, to obtain the coarsely matched images $I^{Ref}_{matched}$ and features $F^{Ref}_{matched}$.

Such an easy matching scheme [43, 7, 35] performs stably on searching for similar patches. However, as shown in Fig. 1, even though highly-similar patches are available in I^{Ref}, there usually exists misalignment in orientation and scales. So far the coarsely-warped reference is not robust to rotation and scaling, which may yield blurry outputs by averaging unaligned Ref and LR. Inspired by [3, 14, 8], we propose to estimate patch-wise spatial transformation A to further align all matched patches in $F^{Ref}_{matched}$:

$$A = T(concat(I^{LR}\uparrow, I^{Ref}_{matched})). \tag{2}$$

Instead of predicting a global transformation like [14] for the whole $F^{Ref}_{matched}$, the local spatial transformer network T is designed to estimate patch-wise alignment parameters for all patches. Each patch of $F^{Ref}_{matched}$ is then aligned independently with the estimated affine matrix to get the fine-aligned reference features $F^{Ref}_{aligned}$. $F^{Ref}_{aligned}$ will be used to facilitate the I^{SR} generation by feature fusion.

3.2. Adaptive Feature Fusion

A direct feature fusion (e.g. concatenation, summation) fails to consider the quality of the matches, which can inevitably bring irrelevant or noisy information. Prior work [35] thus adopted the confidence map as a guidance for feature fusion. But in the original confidence map, C_i is calculated independently for each patch i, which means it reflects the local matching confidence of every single patch, and the transition among neighbor patches is not necessarily smooth. To solve this issue, we embed the confidence map with an extra convolution net g. It is a simple and effective way to aggregate neighbor confidences for more consistent and higher-quality results. The feature fusion process can be represented as:

$$F^{fused} = g(C) \cdot h(F^{SR}, F^{Ref}_{aligned}) + F^{SR}, \tag{3}$$

where $g(\cdot)$ and $h(\cdot)$ are learnable convolution layers.

Another issue is that the images reconstructed from the fused features tend to lose the high-frequency details. Inspired by recent work [36], which generates high-frequency details by adding back image residuals with attention maps, we also conduct adaptive fusion in the image space. The aligned high-frequency (HF) residuals can be represented by $I^{HF}_{aligned} = (I^{Ref} - I^{Ref}\downarrow\uparrow)_{aligned}$. However, different from inpainting task [36] where the HF details have no constraints in the missing regions, in super resolution the details need to be consistent with the original LR contents. To avoid introducing high-frequency noise, we also use a learnable function g_r on the final fusion:

$$I^{SR} = g_r(C) \cdot I^{HF}_{aligned} + decoder(F^{SR}). \tag{4}$$

3.3. Loss Function

We generate the output image conditioned on I^{Ref}, and expect I^{SR} to approximate the ground-truth I^{HR}. Due to misalignment between I^{HR} and I^{Ref}, we found that using I^{HR} as strict labels for the supervised learning leads to unsatisfactory details. We thus adopt the reconstruction term proposed in [21, 22], which calculates losses in low-frequency and high-frequency bands separately :

$$\mathcal{L}_{rec} = \left\| I^{SR}_{blur} - I^{HR}_{blur} \right\| + \sum_i \delta_i(I^{SR}, I^{HR}), \tag{5}$$

where I_{blur} is filtered by 3×3 Gaussian kernels with $\sigma = 0.5$. $\delta_i(X, Y) = \min_j \mathbb{D}_{x_i, y_j}$ is distance between SR pixel x_i and its most similar HR pixel y_i under certain distance $\mathbb{D}$ [22, 40]. The first term softly makes I^{SR} keep the same content as I^{HR} in low-frequency domain. The second term flexibly enforces the statistics of I^{SR} similar to I^{HR}.

We find only using aforementioned losses yields blurry results, as the losses do not involve constrains on I^{Ref}. Intuitively, the fusion modules in Fig 3 (b) can easily ignore the reference information, and degrade to an identity mapping. In this case, I^{Ref} contributes less to I^{SR} generation. To avoid the 'idleness' of I^{Ref}, we introduce a fidelity term modified from [22], where δ_i is the distance between I^{SR} and nearest-neighbor pixels in I^{Ref} under distance $\mathbb{D}$:

$$\mathcal{L}_{fid} = \frac{\sum_i \delta_i(I^{SR}, I^{Ref}) \cdot c_i}{\sum_i c_i}. \quad (6)$$

Pixels with higher matching confidence c_i are given larger weights for optimization, since these pixels can find highly-related cues in I^{Ref}. This fidelity loss can adaptively maximize the similarity between I^{SR} and I^{Ref}. The overall loss is the weighted sum of $\mathcal{L}_{rec}$ and $\mathcal{L}_{fid}$.

3.4. Self-supervised Real-image Adaptation (SRA)

For the real-world DCSR in Fig. 1, we take the wide-angle image I^{wide} and telephoto image I^{tele} as I^{LR} and I^{Ref}, respectively. We aim at super-resolving I^{wide} to produce I^{SR}, but the ground-truth I^{HR} is unavailable to calculate the aforementioned losses. A typical training setting is to downscale the original I^{wide} and I^{tele} by half to simulate the training inputs, and regard original I^{wide} as I^{HR} for supervised learning. However, we found that models trained on downsampled images show significant performance drop on the real images (original I^{wide} and I^{tele}) due to the domain gap between downsampled and real-world images. To bridge this gap, inspired by recent works [1, 30, 31], we propose a self-supervised real-image adaptation strategy (SRA) to fine-tune the trained model M with real-world inputs without ground-truth. Specifically, we directly take the original I^{wide} and I^{tele} from the training set as I^{LR} and I^{Ref}, and the training loss is defined as:

$$\mathcal{L} = \left\| I^{SR} \downarrow - I^{wide} \right\| + \lambda \mathcal{L}_{fid}(I^{SR}, I^{tele}) \quad (7)$$

The first term enforces I^{SR} to preserve the content of I^{wide}, while the second term is to transfer I^{tele} details. After this training stage, the model $M' = \min_M \mathcal{L}$ generalizes well to the real-world inputs.

4. Experiments

4.1. Datasets

CUFED5 [43] It contains 11,871 training pairs and 126 test images. Each test image is accompanied with four reference images ranked by the similarity levels. The resolutions of HR and Ref are about 300×500.

CameraFusion We construct a new dataset for dual-camera super-resolution, which contains 146 pairs of 4k wide-angle and telephoto images in diverse outdoor and indoor scenes. As show in Fig. 1, they share the same scene but differ in ISP and view-point. A compelling RefSR approach is expected to show significant advantages over SISR methods in the overlapped FoV area, while achieving comparable or better performance otherwise.

SISR	PSNR	SSIM	RefSR	PSNR	SSIM
SRCNN [9]	25.33	0.745	Landmark [38]	24.91	0.718
MDSR [18]	25.93	0.777	CrossNet [45]	25.48	0.764
RDN [42]	25.95	0.769	SRNTT [43]	25.61	0.764
RCAN [41]	26.06	0.769	SRNTT-ℓ_2 [43]	26.24	0.784
LapSRN [16]	24.92	0.730	SSEN[28]	26.78	0.791
SRGAN [17]	24.40	0.702	FRM[34]	24.24	0.724
ENet [27]	24.24	0.695	TTSR [35]	25.53	0.765
ESRGAN [32]	21.90	0.633	TTSR-ℓ_1 [35]	27.09	0.804
RSRGAN [39]	22.31	0.635	Ours	25.39	0.733
CSNLN [23]	24.73	0.743	**Ours-ℓ_1**	**27.30**	**0.807**

Table 1. Quantitative comparisons on CUFED5.

SISR	PSNR	SSIM	RefSR	PSNR	SSIM
Bicubic	33.20	0.893	TTSR [35]	35.48	0.915
RSRGAN [39]	33.51	0.873	TTSR-ℓ_1 [35]	36.28	0.928
RCAN [41]	33.94	0.911	Ours	34.41	0.904
CSNLN [23]	36.10	0.927	**Ours-ℓ_1**	**36.98**	**0.933**

Table 2. Quantitative comparison on the CameraFusion dataset.

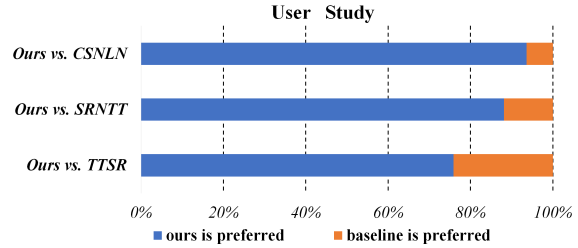

Figure 4. User study results on CUFED5. The reported values indicate the preference rate of our results against other approaches.

4.2. Evaluation

4.2.1 Evaluation on CUFED5

Quantitative Comparison Table 1 shows quantitative comparisons on CUFED5 in terms of PSNR and SSIM. It has been verified that due to trade-off between perception and distortion for super-resolution [33], visually-better results may suffer performance drop of PSNR. Therefore, we follow the setting in previous work [43, 35] to re-train our model with ℓ_1 loss only for fair comparison.

Qualitative Comparison As shown in Fig. 5, our method shows better visual quality on faces, text, objects and textures. In the first example, human faces show with different orientations in the I^{LR} and I^{Ref}, while in the last example, the cruise ship shows a larger size in I^{Ref} than I^{LR} as it is moving forward to the camera. Despite the misalignment of orientations and scales, our model successfully obtains high-fidelity results via robust feature warping and fusion, while other methods either generate abrupt artifacts

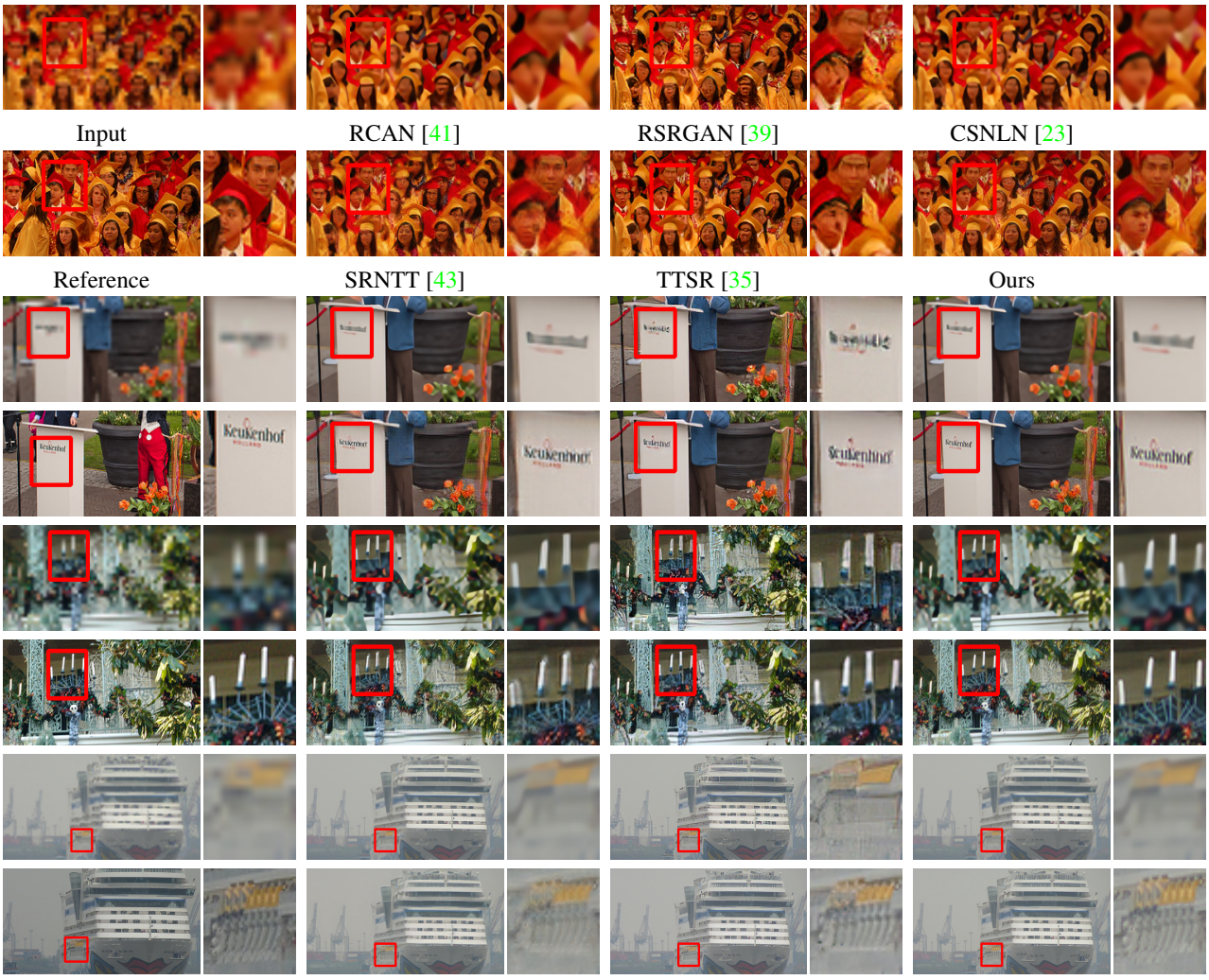

Figure 5. Qualitative comparisons on CUFED5. Our method reconstructs sharper and more realistic details than existing approaches for faces, texts, objects and textures. Zoom-in for details.

or produce blurry details. In other two examples, the scenes are statistic but have different view-points in I^{LR} and I^{Ref}, and we reconstruct recognizable texts and realistic textures.

User Study We conduct a user study on Amazon Mechanical Turk (AMT) to compare our approach with state-of-the-art SISR [23] and RefSR [43, 35] methods. In specific, we provide participants with two images (ours and baselines) each time and ask them to select a more realistic one. We totally collect 1,920 valid votes from 16 participants. As shown in Fig. 4, we outperforms previous work by a large margin.

4.2.2 Evaluation on CameraFusion

To evaluate our method on dual-camera super-resolution, we re-train our model and baselines on the CameraFusion dataset. We select TTSR, CSNLN, RSRGAN and RCAN for comparison considering their outstanding performance on CUFED5. Specifically, we downsample 4K wide-angle and telephoto pairs to 2K-resolution for training and metrics calculation in Table 2, by regarding 4K wide-angle images as ground truth. We also observe that our performance gap between the overlapped FoV (37.28 / 0.942) and other regions (36.94 / 0.931) is small. This implies our approach is robust to reference image with different similarity levels.

For qualitative comparison, we fine-tune the trained models with full-resolution inputs, as mentioned in Section. 3.4. When inference, we can super-resolve the 4K-resolution inputs to obtain 8K results. As in Fig. 6, our approach correctly transfers correlated patterns to reconstruct higher-fidelity outputs within overlapped FoV. It also achieves comparable or better performance outside the overlapped FoV where corresponding reference patches are unavailable.

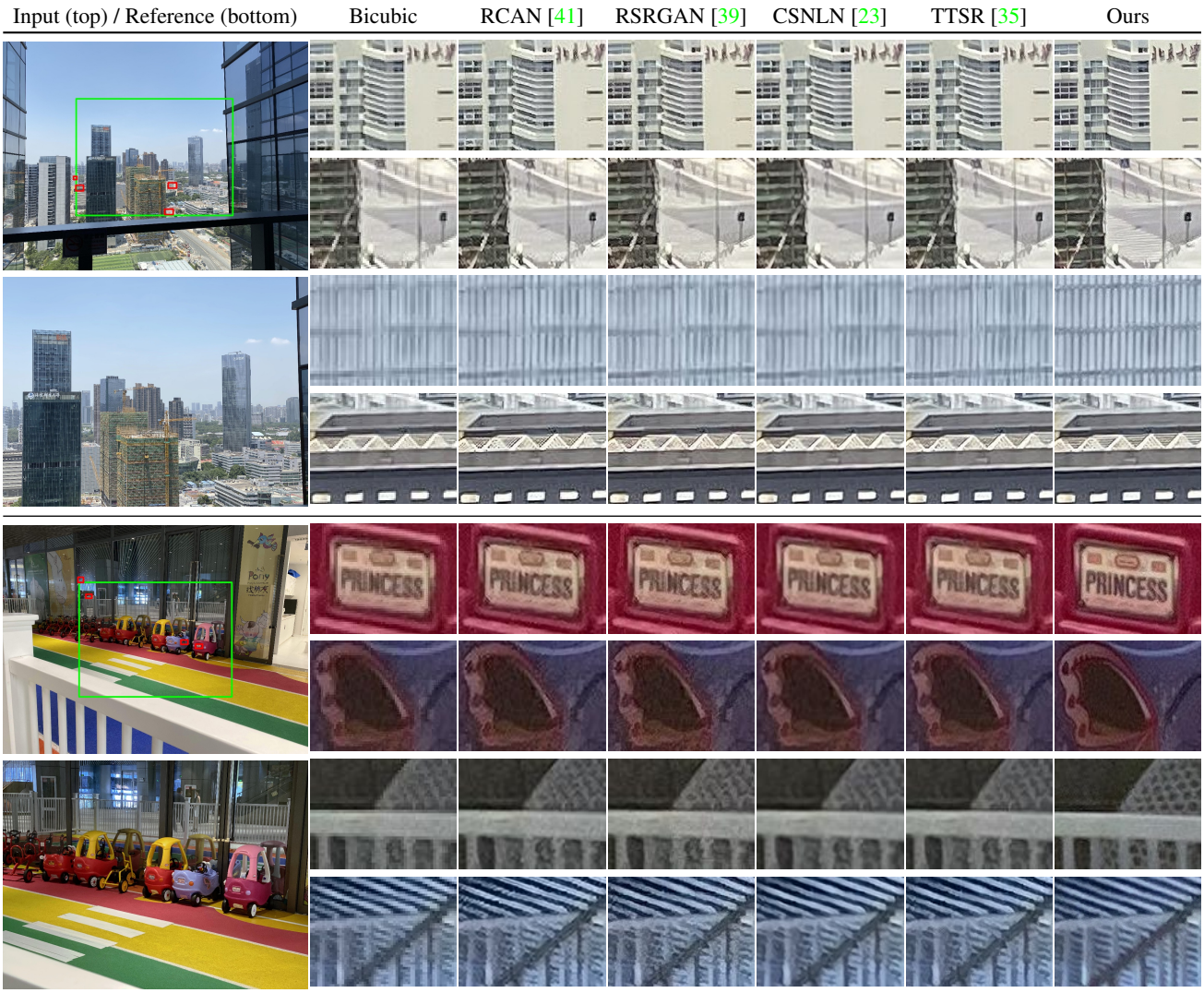

Figure 6. Qualitative comparisons on the CameraFusion dataset. The green box indicates the overlapped FoV area between Input and Ref. Our method reconstructs sharper and more realistic details than previous approaches. Zoom-in for details.

4.3. Ablation Study

4.3.1 Effect of reference similarity-levels

To analyze how the performance of our method is related to the reference images, we conduct experiments on reference images with different similarity levels in CUFED5 [43]. In Table 3, L1 provides the most similar reference images, while L4 is the least relevant level. Our model suffers little degradation when the similarity level decreases, which means that our method can robustly reconstruct images with reference images of different similarity levels.

Similarity level	L1	L2	L3	L4
CrossNet [45]	25.48 / .764	25.48 / .764	25.47 /. 763	25.46 / .763
SRNTT-ℓ_2 [43]	26.15 / .781	26.04 / .776	25.98 / .775	25.95 / .774
SSEN [28]	26.78 / .791	26.52 / .783	26.48 / .782	26.42 / .781
TTSR-ℓ_1 [35]	26.99 / .800	26.74 / .791	26.64 / .788	26.58 / .787
Ours-ℓ_1	**27.30 / .807**	**26.92 / .795**	**26.80 / .791**	**26.70 / .788**

Table 3. Ablation result on the similarity level of reference images. CUFED5 provides four reference images for each LR image ranked by the similarity level, where L1 is the most relevant one.

4.3.2 Effect of Aligned Attention

To further demonstrate how the aligned attention facilitates the feature warping, we directly apply the feature-space index maps learned with and without the aligned attention to warp the original reference image. Note that since the index maps are originally learned to warp feature maps (instead of the images), the warped images are not the SR outputs and only used for visualization. We also visualize the warped reference image by flow-based method [45] and

Feature Fusion Method	PSNR	SSIM
Element-wise Summation	26.85	0.794
Soft Fusion [35]	27.12	0.803
Adaptive Fusion	27.30	0.807

Table 4. Ablation study on different feature fusion methods on CUFED5.

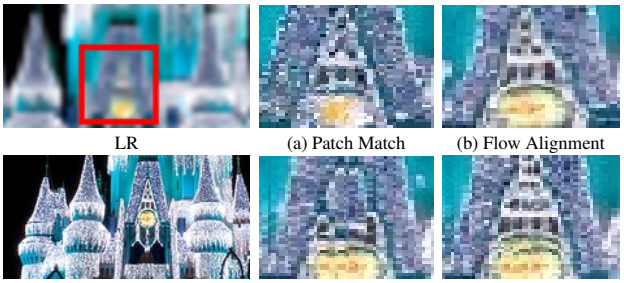

Figure 7. Ablation study on the aligned attention. The building presents different size and viewpoint in input and reference image, and we warp the reference by different methods.

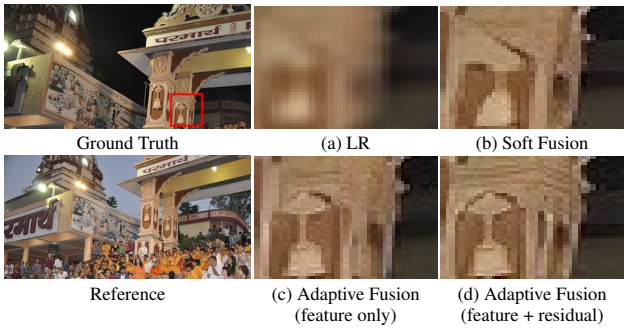

Figure 8. Ablation study on the adaptive fusion. As shown in (c), with adaptive fusion only in the feature space, high-frequency details are not be fully transferred.

Patch Match [2] for comparison. As shown in Fig. 7, flow-based alignment leads to distorted structures, while Patch Match lacks high-quality details. In contrast, our model can alleviate the spatial misalignment.

4.3.3 Effect of Adaptive Fusion

Table 4 provides ablation results on the adaptive fusion. We apply element-wise summation (add $F^{ref}_{aligned}$ to F^{SR} without confidence guidance), soft fusion [35] (fuse $F^{ref}_{aligned}$ and F^{SR} with original confidence map) and adaptive fusion (fuse $F^{ref}_{aligned}$ and F^{SR} with learnable confidence map), respectively. With the adaptive fusion, we observe performance gain of 0.18 dB over the soft fusion, which implies the benefit from a learnable confidence map. As shown in Fig. 8, by further applying adaptive fusion for high-frequency residuals as Eq. 4, the model can generate sharper structures and more realistic textures.

Figure 9. Ablation experiment on the fidelity loss. With the fidelity loss, we can obtain higher-fidelity reconstruction results.

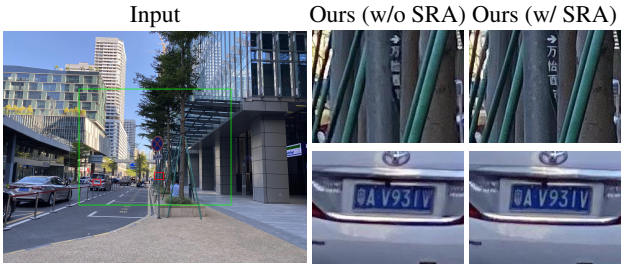

Figure 10. Ablation study on SRA on CameraFusion dataset. Zoom-in for details.

4.3.4 Effect of Fidelity Loss

The fidelity loss is imposed to enforce the output SR image to possess high-quality details as reference images. The key idea is to adaptively maximize the similarity between Ref and SR according to the matching confidence. Fig. 9 shows that without this loss, the network fails to accurately utilize reference cues for high-fidelity generation, since LR features dominates the reconstruction process.

4.3.5 Effect of Self-supervised Real-image Adaption

As shwon in Fig. 10, without the proposed self-supervised real-image adaption, the super-resolution results on real-world camera photos are blurry.

5. Conclusion

In this paper, we study the reference-based super-resolution with the focus on real-world dual-camera zoom. To alleviate the spatial misalignment between input and reference images, we propose an aligned attention module for more robust feature warping. To advance our method to dual-camera super-resolution for real-world smartphone images, we design a self-supervised domain adaptation scheme to generalize trained models to real-world inputs. Extensive experiments show that our method achieves compelling performance.

Acknowledgement

This project is supported by SenseTime Collaborative Research Grant. We thank Hao Ouyang and anonymous reviewers for helpful discussions and suggestions.

References

[1] Michal Irani Assaf Shocher, Nadav Cohen. "zero-shot" super-resolution using deep internal learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2, 8

[3] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *European Conference on Computer Vision*, pages 29–43. Springer, 2010. 3, 4

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 3

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3086–3095, 2019. 1, 3

[6] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 2020. 1

[7] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 3, 4

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 4

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2, 5

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[11] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[13] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1604–1613, 2019. 1, 3

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 4

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[16] 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. 5

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan 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*, pages 4681–4690, 2017. 2, 5

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2, 5

[19] Yucheng Liu and Buyue Zhang. Photometric alignment for surround view camera system. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1827–1831. IEEE, 2014. 3

[20] Sai Kumar Reddy Manne, BH Pawan Prasad, and KS Green Rosh. Asymmetric wide tele camera fusion for high fidelity digital zoom. In *International Conference on Computer Vision and Image Processing*, pages 39–50. Springer, 2019. 3

[21] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conference on Computer Vision*, pages 427–443. Springer, 2018. 4

[22] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 4, 5

[23] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 6, 7

[24] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009. 3

[25] Seoyoung Park, Byeongho Moon, Seonhee Park, Seungyong Ko, Soohwan Yu, and Joonki Paik. Brightness and color correction for dual camera image registration. In *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–2. IEEE, 2016. 3

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3

[27] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through

automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017. 5

[28] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2020. 1, 2, 5, 7

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9446–9454, 2018. 5

[31] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image inpainting with external-internal learning and monochromic bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5120–5129, 2021. 5

[32] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2, 5

[33] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[34] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *European Conference on Computer Vision*, pages 230–245. Springer, 2020. 1, 2, 5

[35] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[36] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 4

[37] Soohwan Yu, Byeongho Moon, Donggyun Kim, Sehoon Kim, Wonhee Choe, Sangkeun Lee, and Joonki Paik. Continuous digital zooming of asymmetric dual camera images using registration and variational image restoration. *Multidimensional Systems and Signal Processing*, 29(4):1959–1987, 2018. 3

[38] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu. Landmark image super-resolution by retrieving web images. *IEEE Transactions on Image Processing*, 22(12):4865–4878, 2013. 5

[39] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3096–3105, 2019. 2, 5, 6, 7

[40] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3762–3770, 2019. 1, 4

[41] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2, 5, 6, 7

[42] 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. 5

[43] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019. 1, 2, 3, 4, 5, 6, 7

[44] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, Haoqian Wang, Yebin Liu, and Lu Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In *BMVC*, 2017. 1

[45] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu 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)*, pages 88–104, 2018. 1, 2, 5, 7

[46] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984. IEEE, 2011. 1