Geometry Enhanced Reference-based Image Super-resolution

Han Zou[1,2] Liang Xu[1] Takayuki Okatani[1,2]
[1]Graduate School of Information Sciences, Tohoku University [2]RIKEN Center for AIP
{hzou, xu, okatani}@vision.is.tohoku.ac.jp

Abstract

With the prevalence of smartphones equipped with a multi-camera system comprising multiple cameras with different field-of-view (FoVs), images captured by two or three cameras now share a portion of the FoV that are compatible with reference-based super-resolution (RefSR). In this work, we propose a novel RefSR model that utilizes geometric matching methods to enhance its performance in two aspects. First, we integrate geometric matching maps to improve feature fusion. Second, we train the matching modules equipped in the RefSR models under the supervision of accurate geometric matching maps to increase their robustness. Our experimental results demonstrate the effectiveness and state-of-the-art performance of the proposed method.

1. Introduction

Some of the recent mobile phones are furnished with multiple cameras with the purpose of augmenting the overall photographic experience. The included cameras possess different focal lengths, enabling users to capture photos and videos with divergent fields of view (FoVs) in unison. For instance, when two cameras with different FoVs capture the same scene, the camera with a more restricted FoV generates a higher-resolution image within the overlapped FoV. The incorporation of this image with the same scene image captured by the other camera with a wider FoV can facilitate an increase in image resolution.

The problem of enhancing the resolution of a low-resolution (LR) input of a scene with the assistance of a reference (Ref) image of a similar scene is referred to as a reference-based image super-resolution (RefSR). Researchers have recently explored this problem, resulting in the development of several RefSR methods [5,10,28,29,32–34]. These methods function by initially establishing correspondences between the LR and Ref images, followed by warping the Ref image to align it with the LR image, and finally fusing the two images to generate high-quality SR images. The LR and Ref image pairs obtained via a multi-camera system can be regarded as a specific RefSR configuration, in which the LR and Ref images share a portion of the field of view (FoV).

The primary challenge of RefSR is obtaining accurate correspondences between low-resolution (LR) and reference (Ref) images. Previous works have typically employed the patch matching method [1] to accomplish this objective. In particular, they initially extract features of patches at each position of LR and Ref images, and then establish correspondences between LR and Ref images by selecting patch pairs possessing the highest similarity. Such patch matching is implemented as a sub-module of the entire network for RefSR. To enhance the efficacy of the matching module in producing high-quality SR outcomes, it is typically trained via a supervised approach incorporating a reconstruction loss evaluating the alignment of the two images such as L1 loss. Despite the success achieved through this design, there exists considerable room for further improvement.

Firstly, the patch-matching method is so flexible and can exhibit the capability to establish correspondences between LR and Ref image pairs on all positions. However, they do not incorporate geometric constraints pertaining to the LR and Ref images. It is important to note that a geometric transformation of the Ref image can enable its coarse alignment with the shared region of the LR image. Because geometric constraints are not utilized during the training process, the matching module tends to generate matching maps that are incongruent, even in the shared regions. Thus, there is an opportunity to enhance the reliability of Ref feature matching and fusion by integrating geometric matching into the aforementioned methodology, with the aim of generating features that possess both smoothness and sharpness.

Secondly, although a portion of the scene is shared between the LR and Ref images, which enables the successful establishment of correspondence between features within the shared region, matching features in the unshared region still presents a significant challenge. This issue becomes more significant when training the model on real-world images because the absence of a ground truth hinders the supervision of SR generation. Previous works [10, 28]

have adopted a self-supervised approach to train the RefSR model on real-world datasets. Specifically, they leverage images possessing narrower FoVs, such as telephoto images, as a pseudo-ground truth and minimize the texture differences [12] between the SR output and these images. However, this pseudo ground truth is not aligned with the SR images, leading to considerable geometric misalignment and color differences. Consequently, the texture difference may not be reliable, causing the model to learn to generate false textures, particularly outside the shared region. As the matching module is trained alongside the entire framework, it may be misled to produce false matches during back-propagation, particularly for the matching module between ultra-wide and telephoto images that share only a small region of the scene.

To address the aforementioned issues, we leverage pre-trained geometric matching methods to enhance the RefSR model's performance. Geometric matching methods are capable of accurately estimating the geometric transformation between the LR and Ref images. We mainly utilize the estimated maps in two ways to enhance the RefSR performance. Firstly, we concurrently utilize both geometric and patch matching maps to warp the Ref features. The use of geometric matching maps enables the warped Ref features to possess greater smoothness within the shared region. Furthermore, the independent matching module can reduce the artifacts caused by the mismatching when deploying the entire model to real-world images.

The second approach we employ to enhance the RefSR model is to improve the matching robustness by utilizing geometric matching maps. Prior attempts to develop an effective patch matching method that can estimate matching maps between ultra-wide and telephoto images have been unsuccessful due to the significant difference in FoVs. However, geometric matching methods can be employed to obtain relatively accurate matching maps between them. The estimated maps can then be employed as pseudo ground truth to facilitate the training of patch matching method. To achieve this, we adopt the margin triplet rank loss [4], which is commonly utilized in image matching tasks, and train patch matching method on pairs of real-world images, such as ultra-wide and telephoto images.

The integration of the geometric matching method has proven to be successful in enhancing the RefSR performance on datasets collected from multi-camera systems. Additionally, by utilizing the patch matching method strengthened with geometric information, we have mitigated artifacts resulting from mismatching and achieved an improvement in the visual quality of the real-world SR results.

2. Related Work

2.1. Reference-based Super-Resolution

Reference-based super-resolution (RefSR) uses additional reference image(s) to super-resolve an input low-resolution image. Previous studies have shown the effectiveness of transferring information from a high-resolution reference image to generate SR images. A critical problem is accurately aligning the Ref image with the LR image, which is important for fusing their image features in a subsequent step to generate high-quality SR images. Zheng et al. [34] estimate optical flows between them for their alignment. Zhang et al. [32] propose using patch matching [1], and Yang et al. [30] improve it by adopting attention mechanisms for feature fusion. Jiang et al. [8] integrate patch matching with modulated deformable convolution, and have subsequently enhance the matching robustness using a knowledge distillation method. Wang et al. [28] propose an aligned attention method for better feature fusion, which preserves high-frequency features via spatial alignment operations well. Huang et al. [5] decouple ref-based SR task into two sub-tasks: single image SR task and texture transfer task, and train them independently. It reduces misuse and underuse of the Ref feature, which often happens in Ref feature transfer. Lee et al. [10] propose RefVSR, which integrates RefSR with VSR.

2.2. Dense Geometric Matching

Dense geometric matching aims to establish a dense pixel-wise correspondence between image pairs under a geometric transformation. Some of previous works aim at acquiring correspondences between instances that pertain to the same class within a semantic plane. [2, 9, 14–16, 18]. To effectively manage the significant displacement resulting from geometric transformations, the majority of recent works rely on dense flow regression approaches, which have been widely adopted in the optical flow methods [3,6,7,19–22]. Melekhov et al. [13] proposed DGC-Net, which employs a CNN-based approach to generate dense correspondences in a coarse-to-fine manner. Rocco et al. [17] proposed a neighborhood consensus network that can filter out ambiguous correspondences by considering the similarities of neighboring matches and enforcing a consensus among them. Truong et al. [24] introduced GLU-Net, a method that combines global and local correlation layers to estimate dense geometric correspondence without restricting the input resolution. Moreover, Truong et al. further enhanced this network's performance by devising an online optimization approach that performs replacements in the feature correlation layers [23]. Jiang et al. [8] utilized both dense and sparse methods and proposed a transformer-based architecture that recursively operates at multiple scales to achieve accurate correspondence estimation. Truong et al. [26, 27]

leverage warp consistency constraints to achieve enhanced unsupervised learning of dense matching. These methods have shown promising results in various applications and provide a valuable contribution to the field of dense geometric correspondence estimation.

3. Analysis on Feature Matching

In RefSR, we consider a low-resolution input image $I^{LR}(\in \mathbb{R}^{H\times W\times C})$ and a reference input image $I^{Ref}(\in \mathbb{R}^{H\times W\times C})$, where H, W, and C denote the height, width, and number of channels of the input images, respectively. Here, I^{LR} is an image with a large field of view (FoV) captured at a low resolution, while I^{Ref} is an image of the same scene with a narrower FoV captured at a higher resolution (e.g., using wide and telephoto lenses). Our objective is to generate a super-resolved image $I^{SR}(\in \mathbb{R}^{sH\times sW\times C})$ of I^{LR} with an upscaling factor s.

Recent works utilize the reference image to improve the quality of the super-resolved output by estimating pixel-wise correspondences between the input image pair. Specifically, high-resolution features from the reference image are warped using the estimated correspondences to refine the low-resolution features. Two types of matching models can be used to accomplish this objective, namely patch matching and geometric matching. The former is a prevalent method in recent RefSR studies.

3.1. Patch Matching Method

Methods based on patch matching establish correspondence between extracted local feature patches from input image pairs. The method first embeds the input image pair into feature maps f_p^{LR} and f_p^{Ref} using a feature encoder ϕ_1, and densely extracts 3×3 patches with a stride of 1. Next, it computes the cosine similarity $S(i, j)$ between all pairs of LR feature patches and Ref feature patches, where i, j are the indexes of LR and Ref feature patches. In order to derive the matching map, the method selects the patch with the highest cosine similarity from all the Ref feature patches, and designates it as the matched patch of a LR feature patch. The cosine similarity score associated with this matched patch is used as the confidence score. The matching map and its confidence map can be formulated as follows:

$$M_p(i) = \arg\max_j S(i, j), \tag{1a}$$

$$C_p(i) = \max_j S(i, j). \tag{1b}$$

The resulting matching map M_p is formed by aggregating all the matched indexes from the Ref feature map to the LR feature map, while the confidence map C_p reflects the reliability of the matched features. The aligned Ref feature is then obtained by warping Ref features with the estimated index map, which is subsequently used to refine the LR feature. This method is typically trained together with the entire RefSR model and supervised by an image reconstruction loss.

3.2. Geometric matching Method

Geometric matching methods are designed to estimate geometric displacements between pairs of images. Similar to patch matching methods, these methods typically rely on estimating image correspondences or optical flow by computing local similarities in the feature space. Recent works have leveraged both global and local correlations [24] to achieve precise correspondence across all locations in the feature map pairs. Notably, the fundamental distinction from patch matching methods lies in the training of the geometric matching model on a dataset that adheres to a geometric constraint, whereby the estimated correspondences are subject to the constraints of a specific geometric transformation. This constraint limits correspondence solely to the content present in both the LR and Ref images. We denote the matching maps estimated by geometric matching methods as M_g.

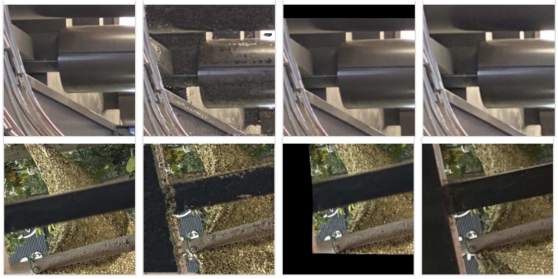


Figure 1. Warped Ref images using patch and geometric matching methods. Specifically, the Ref and LR images are input into two matching methods, after which the Ref image is warped based on the estimated maps in order to align with the LR image.

3.3. Comparison of Two Methods

Both patch matching and geometric matching methods share a similar fundamental principle, albeit with different impacts on establishing feature correspondence, owing to their distinct training objectives. Figure 1 presents the warped Ref features, where RGB images are employed for enhanced visualization, by utilizing correspondences estimated by different matching methods. Patch matching methods facilitate the matching of feature patches without additional constraints, thereby enabling the flexible matching of features from any image region. However, this approach inevitably generates mismatched patches, and numerous inconsistent patches appear in the warped features,

as it is not restricted to a specific geometric transformation. On the other hand, geometric matching methods are trained on datasets that pertain to multi-view or optical flow estimation. By constraining the output correspondence to a specific geometric transformation, consistent transformation results can be generated. The warped results of geometric matching estimation methods, as depicted in Fig 1, treat the image holistically and accurately determine the overlapped region's location in the LR image. However, the correlations cannot be intelligently established beyond the overlapped region, even if the Ref image's content is similar.

4. Proposed Method

The two matching methods have their own strengths and weaknesses, as shown above. In order to establish improved correspondence between images possessing differing FoVs, a potential solution is to integrate these two approaches. Accordingly, we present a novel method for enhancing the performance of RefSR through the utilization of geometric matching methods. This is mainly accomplished in two ways: firstly, by fusing Ref features warped using flows estimated by both geometric and patch matching methods, thereby generating better aligned Ref features. To achieve this, we propose a neural module named the Dual Alignment and Aggregation. Secondly, we enhance the robustness of the patch-based matching method by utilizing the matching outcomes of the geometric matching method. An overview of our proposed method is presented in Fig. 2, wherein the left-hand part depicts the overall framework consisting of three distinct matchers. The Patch matcher is a conventional patch matching module employed in previous studies, while the Geometric matcher is a pre-trained geometric matching model. The Geo-Enhanced matcher is a patch matching module enhanced by geometric matching maps, typically trained for matching features between ultra-wide and telephoto images with large FoV differences.

The right part of Fig.2 illustrates the approach for utilizing a geometric matcher to enhance the robustness of the patch matcher. The detail of this process is provided in Sec.4.2.

4.1. Dual Alignment and Aggregation Module

The input LR and Ref images are denoted by I^{LR} and I^{Ref}, respectively. To generate correspondence maps using the two matching methods, we initially provide I^{LR} and I^{Ref} as input to both a patch matcher and a geometric matcher separately. As described in Sec. 3, the patch matcher outputs the matching map and confidence map M_p and C_p, respectively, while the geometric matcher outputs the matching map M_g. Both of these matching maps can be interpreted as the flow map between I^{LR} and I^{Ref}. The next step is to warp the Ref features f^{Ref} using these two maps, resulting in two warped Ref features.

$$f_p^{Ref} = \mathcal{W}(f^{Ref}, M_p), \tag{2a}$$
$$f_g^{Ref} = \mathcal{W}(f^{Ref}, M_g), \tag{2b}$$

The suitability of the warped features is suboptimal due to discrepancies in resolution, misalignment, and mismatching. In accordance with the methods used in previous research endeavors [8, 28], we employ deformable convolution layers to achieve further alignment of the warped reference (Ref) features. By utilizing deformable alignment, we can introduce offset diversity to enhance the quality of the warped textures. In contrast to prior studies, we contend with Ref features that have been warped from two distinct flows. In order to maximize the benefits of these two matching flows, we conduct simultaneous alignment of the Ref features. Specifically, we concatenate the two warped Ref features and calculate their corresponding offsets. We subsequently apply a deformable convolution network (DCN) to generate fusion results.

$$o_{p,g} = c_o(f^{LR}, f_{p,g}^{Ref}) + M_{p,g}, \tag{3a}$$
$$m_{p,g} = \sigma(c_m(f^{LR}, f_{p,g}^{Ref})), \tag{3b}$$
$$f_{fusion}^{Ref} = DCN(f^{Ref}, o_{p,g}, m_{p,g}), \tag{3c}$$

where c^o and c^m denote stacks of convolutional layers for generating offset and mask, and σ is the sigmoid function. f_{fusion}^{Ref} is the resulting aligned Ref feature of two matching methods. Directly fusing aligned features with LR frame features may bring unreliable information due to error-prone matching results [28]. So we first follow prior works to adaptively select Ref features from aligned Ref features. We take confidence map C_p to guide the Ref feature fusion.

$$\hat{f}_p^{SR} = g_1(C_p) \cdot g_2(f^{LR}, f_{fusion}^{Ref}), \tag{4}$$

where g_1 and g_2 indicates two conv layers. Next, Ref features aligned from different matching methods affect different regions of the image.

4.2. Geometry-Enhanced Matching

The approach described in a previous study [10] was unable to effectively transfer the high-resolution features of the telephoto image I^{Tele} to the ultra-wide image I^{UW}. This is due in part to the substantial difference in FoV between these two images, making it challenging to accurately match features. Moreover, the unavailability of 8K ground truth images exacerbates the challenges involved in training the matching module in conjunction with the entire RefSR model through a self-supervised approach. As discussed in Section 3, matching methods based on geometry can produce more polished but less detailed warping flows, as they

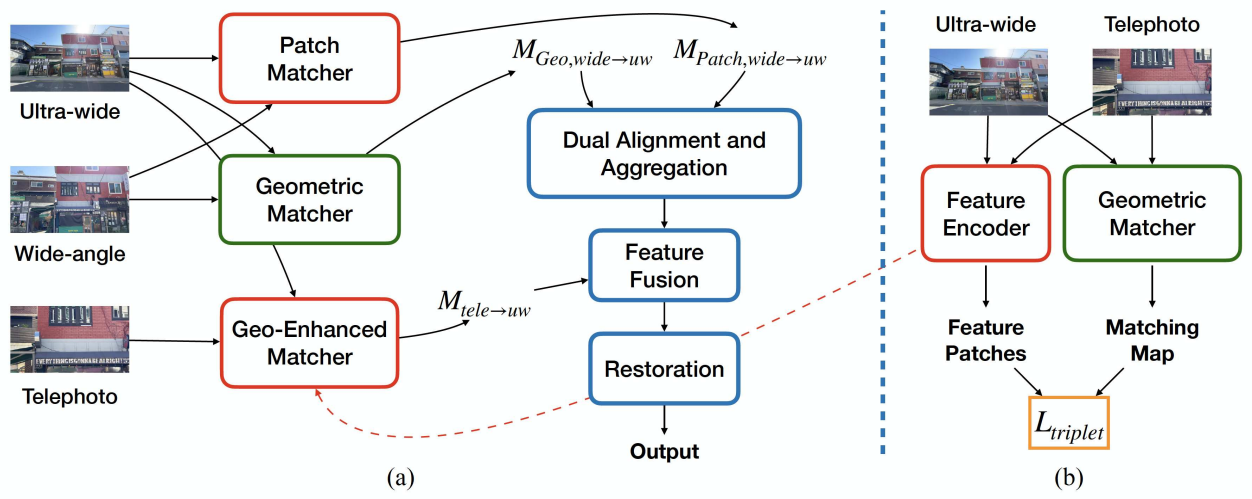


Figure 2. (a) The overview of the proposed method. Three real-world images are fed into the network with the objective of producing an 8K ultra-wide image. The patch matcher and geometric matcher modules estimate two matching maps between the ultra-wide and wide-angle images, which are then passed to the Dual Alignment and Aggregation module. Subsequently, the Geo-Enhanced Matcher is employed to estimate matching maps between the ultra-wide and telephoto images, which are fused with the previously obtained features. Finally, the resulting features are utilized to generate the 8K ultra-wide output. (b) The training process of the Geo-Enhanced Matcher involves the initial estimation of geometric matching maps, which serve as the pseudo ground truth. These maps are subsequently utilized to train the feature encoder, whereby the distance between feature patches at correspondence positions for the matching map is minimized. Following this, the pre-trained feature encoder is employed in the Geo-Enhanced Matcher.

are trained with geometric constraints. One solution is to develop an independent matching module, which can be trained using geometric matcher outcomes as a surrogate for ground truth.

Inspired by [4, 8], we leverage the triplet margin ranking loss to attain the desired objective. We first embed two images I^{Tele} and I^{UW} into feature maps via two encoders and dense extract feature patches. Next, we employ a pre-trained geometric matcher to estimate warping flows between I^{Tele} and I^{UW}, from which we derive the corresponding pseudo pair p_{tele} in I^{Tele} for every position p_{uw} in I^{UW}, which serves as positive examples. To obtain negative samples, we identify the most challenging pairs lying beyond a square local neighborhood of dimension $2K$.

$$N1 = \underset{q \in I^{Tele}}{\arg\min} ||f_{p_{uw}} - f_q||_2, \text{ where } ||q - p_{tele}||_\infty > K, \tag{5a}$$

$$N2 = \underset{q \in I^{UW}}{\arg\min} ||f_{p_{tele}} - f_q||_2, \text{ where } ||q - p_{uw}||_\infty > K, \tag{5b}$$

where f_x denotes the feature patch located at position x. To ensure that the matching module establishes accurate correspondences among positive pairs, our aim is to minimize the distance between feature patches of such pairs, while maximizing the distance between negative pairs. The positive distance and negative distance are subsequently computed from their respective pairs.

$$P(p_{uw}) = ||f_{p_{uw}} - f_{p_{tele}})||_2, \tag{6a}$$

$$N(p_{uw}) = min(||f_{p_{uw}} - f_{N1})||_2, ||f_{p_{tele}} - f_{N2})||_2), \tag{6b}$$

The triplet margin ranking loss is calculated by

$$L_{triplet} = max(0, M + P(p_{uw})^2 - N(p_{uw})^2), \tag{7}$$

By utilizing the aforementioned methods to train the feature encoders, it becomes possible to establish a correspondence between the extracted feature patches of the telephoto and ultra-wide images. These two Geo-Enhanced feature encoders can then be utilized to construct a patch matcher, referred to as the Geo-Enhanced matcher, as shown in the Fig. 2. To compare the efficacy of the patch-matching modules, we evaluated those trained using the proposed methods against those trained concurrently with the RefSR model in its entirety in Sec. 5.6.

After obtaining the estimated matching map and its corresponding confidence from the Geo-Enhanced matcher, the Ref features are fused using the method presented in Eq. 4. The resulting feature is then utilized for the purpose of generating the final SR output.

4.3. Training

The proposed model is trained using a Ref input with a narrower Field of View (FoV) in the presence of an LR input, with the resulting output, denoted as I^{SR}, anticipated

to closely approximate the ground truth image while incorporating more refined texture details. Note that $4\times$ down-sampled ultra-wide and wide-angle images serve as LR and Ref inputs, respectively, and the original ultra-wide images I^{UW} are treated as ground-truth. Previous works [10, 28] have typically employed two distinct losses. The first, a reconstruction loss, is intended to preserve the contents of the LR input, while the second, a reference fidelity loss, is intended to facilitate the transfer of texture details from the Ref inputs. More specifically, the reconstruction loss is utilized to compute the low-frequency and high-frequency bands between I^{SR} and I^{UW}.

$$\ell_{rec} = ||I^{SR}_{blur} - I^{UW}_{blur}|| + \alpha \sum_i \delta_i(I^{SR}, I^{UW}), \quad (8)$$

where I_{blur} indicates images are filtered by 3×3 Gaussian kernels with $\sigma = 0.5$. The contextual loss, defined as $\sum_i \delta_i(X, Y) = min_j \mathbb{D}_{x_i, y_j}$, calculates the distance between the SR pixel x_i and its most comparable HR pixel y_j at a certain distance $\mathbb{D}$. The reference fidelity loss serves as guidance for feature fusion of the Ref images, and is expressed as:

$$\ell_{fid} = \sum_i \delta_i(I^{SR}, I^{Wide}), \quad (9)$$

where I^{Wide} is the original size wide-angle image. The loss is the weighted sum of reconstruction loss and reference fidelity loss:

$$\ell_{tex} = \ell_{rec} + \beta \ell_{fid}, \quad (10)$$

4.4. Adaption to the real-world images

To address the performance degradation associated with training the model on down-sampled input, we follow previous works [10, 28], and undertake fine-tuning of the pre-trained model using real-world images of the original size. Specifically, the original ultra-wide images I^{UW} are utilized as LR inputs, while wide-angle images I^{Wide} and telephoto images I^{Tele} serve as Ref inputs, respectively. Given the lack of ground-truth 8K ultra-wide images, we resort to an approximate loss, whereby the original wide-angle and telephoto images serve as pseudo ground-truths, in a self-supervised manner. The training loss is formulated as follows:

$$\ell_{ada} = ||I^{SR}_{\downarrow,blur} - I^{UW}_{blur}|| + \gamma \ell_{fid}(I^{SR}, I^{Tele}), \quad (11)$$

where γ is a weighting constant.

5. Experiments

5.1. Datasets

We use two datasets in our experiments, the CameraFusion [28] and RealMCVSR [10] datasets.

RealMCVSR [10] consists of 161 video triplets recorded by a triple camera system equipped on iPhone 12 Pro Max. Each video triplet has three videos captured in the same scene but with different FoVs: ultra-wide, wide-angle, and telephoto. The dataset uses HD resolution(1080×1920) and totally have 23107 frames.

CameraFusion [28] consists of 146 pairs of 4k wide-angle and telephoto images with outdoor and indoor scenes. All the image pairs are captured by a dual-camera system. We only use it to evaluate the performance of the proposed Dual Alignment and Aggregation module since it does not have an ultra-wide image.

5.2. Implementation

We use VGG19 pre-trained on ImageNet1K for encoding features in the patch matching module, and PDCNet [25] pre-trained on MegaDepth [11] for generating geometric correspondence between LR and Ref pair. All the geometric correspondence are pre-generated using $4\times$ down-sampled image images. We train the proposed model using Adam optimizer. The learning rate is initialized with 2.0×10^{-4} and steadily decreased to 1.0×10^{-6} using the cosine annealing strategy. The size of the ultra-wide, wide-angle, and telephoto input patches are 64×64 ,128×128, and 256×256, respectively. The loss weights α,β, and γ are set to 0.01, 0.05, and 0.1, respectively.

To train models on images down-sampled by a factor of $4\times$, we employ down-sampled ultra-wide and wide-angle images as LR input and Ref input, respectively. At this stage, the Geo-Enhanced matcher is not utilized, as supervision is provided by the ground truth, and the matching results between wide-angle and ultra-wide images are deemed trustworthy. The matcher is trained concurrently with the entire framework. During the transition to real-world image adaptation, we incorporate matching between ultra-wide and telephoto images. Given the significant FoV gaps between them and the inevitable limitations associated with a self-supervised approach, we leverage the pre-trained Geo-Enhanced matcher to facilitate image matching between them.

5.3. Quantitative Comparison

We quantitatively evaluate the proposed method on the RealMCVSR test set. First, we evaluate methods on the model without real-world adaption, i.e., using $4\times$ down-sampled ultra-wide and wide-angle images as the LR and Ref inputs, respectively.

Table 1 shows the results. We select several SR methods, i.e., Bicubic, RCAN [31], TTSR [30], and DCSR [28]. Bicubic and RCAN do not utilize a reference, while others are reference-based methods, which is indicated by 'R-' in the type column. The methods with '-ℓ_1' in the method column of Table 1 indicate that they are trained with $\ell_1$loss

alone, which are used for a fair comparison with previous works. We can see that our method outperforms all the previous SR methods in each category. Both types of models are trained and evaluated with $4\times$ downsampled ultra-wide and wide-angle frames.

5.4. Qualitative Comparison

We compare the performance of $4\times$ on high-resolution image inputs, and the SR results are 8K images. For RCAN [31], we retrained them using the same training configuration as proposed methods without reference input on RealMCVSR. Fig. 4 shows qualitative comparison.

Model	Ref	PSNR	SSIM
Bicubic	✗	26.65	0.800
RCAN-ℓ_1 [31]	✗	31.07	0.915
TTSR [30]	✓	30.31	0.905
TTSR-ℓ_1 [30]	✓	30.83	0.911
DCSR [28]	✓	30.63	0.895
DCSR-ℓ_1 [28]	✓	32.43	0.933
Ours	✓	31.09	0.912
Ours-ℓ_1	✓	**32.70**	**0.935**

Table 1. Quantitative evaluation on the RealMCVSR dataset

5.5. Ablation Study

We have conducted ablation studies to evaluate the impact of the proposed components on RealMCVSR and CameraFusion dataset. Table 2 shows the PNSR performance on $4\times$ down-sampled images. The first row corresponds to a model equipped with a geometric matcher only, while the second row corresponds to a model equipped with a patch matcher only. The performance of the geometric matcher is inferior to that of the patch matcher, as it only affects a portion of the region. In the third row, we have employed both matchers, and directly fused the Ref features warped using the two types of matching results with a set of convolutional layers. The fourth row corresponds to the model equipped with the proposed dual alignment and aggregation module, which achieves the highest performance. It should be noted that the RealMCVSR dataset employs ultra-wide and wide-angle images as the LR and Ref inputs, whereas the CameraFusion dataset utilizes wide-angle and telephoto images as the LR and Ref inputs.

5.6. Analysis on Geo-Enhanced matching

To gain further insights into the effectiveness of the proposed Geo-Enhanced matching methods, we conduct a comparative analysis between the matcher trained with the entire framework and that trained using geometric information. Specifically, we visualize the warped Ref features in Fig. 3, with RGB images being utilized to enhance visualization. Our analysis of the matching results reveals that the matcher trained under the self-supervised loss produces accurate matching results within the shared FoV of the LR and Ref images. However, it is observed to be less reliable in predicting matching maps outside the shared regions. Conversely, the matcher trained using geometric matching results yields satisfactory matching results both inside and outside the shared regions. Despite the presence of content gaps resulting in mismatches within the matching output, there is considerable potential for enhancing the generation of 8K images.

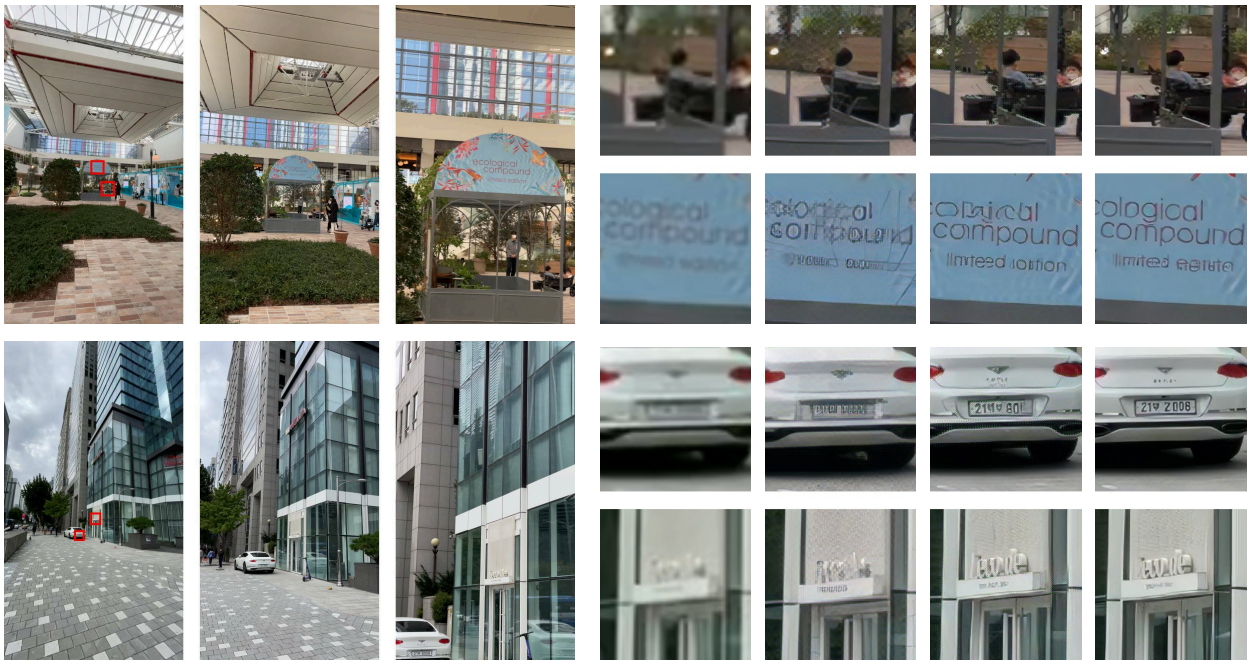
	RealMCVSR		CameraFusion	
	Full	Center	Full	Center
Geo	32.31	34.85	29.95	30.68
Patch	32.51	34.94	30.15	30.75
Patch+Geo	32.63	35.24	30.25	30.99
Patch+Geo+DAA	32.70	35.39	30.37	31.16

Table 2. Abaltion study on the proposed methods. Evaluation(PSNR) is conducted on both full and center-cropped images, where the latter pertains to regions that correspond closely to the overlapping FoV between LR and Ref images.

(a) Input (b) Ref (c) w/o Geo (d) w Geo

Figure 3. Comparison between patch matcher trained with/without geometric information.(a) and (b) are the HR input ultra-wide and telephoto images. (c) and (d) are the warped telephoto images trained with : (c) the whole RefSR model; (d) geometric information.

We proceeded to evaluate the quality of the generated 8K images with and without a Geo-Enhanced matcher. We conduct a comparative analysis of three variations of the model: the first variant being without telephoto input, the second variant incorporating telephoto input but lacking a Geo-Enhanced matcher, and the third variant comprising a



(a) HR inputs (b) Bicubic (c) RCAN [31] (d) DCSR [28] (e) Ours

Figure 4. Quantitative comparison of 8K super-resolution results.

Geo-Enhanced matcher. The corresponding results are presented in Fig. 5. Our findings demonstrate that the utilization of an inadequate matching module in conjunction with telephoto input leads to the degradation of SR generation. Conversely, the integration of a Geo-Enhanced matcher results in the generation of superior letters and numbers, compared to the other two variants. Furthermore, our proposed method effectively mitigates artifacts stemming from mismatches.

(a) HR input (b) Bicubic (c) w/o tele (d) w/o Geo (e) w Geo

Figure 5. Qualitative comparison on HR inputs with/without telephoto inputs and with/without Geo-Enhanced matcher.

6. Summary and Conclusion

In this study, we have presented a new method for reference-based super-resolution. Our approach integrates both patch matching and geometric matching methods, with particular attention paid to the often overlooked usefulness of geometric constraints between images. In order to maximize the efficacy of both matching methods, we have developed a Dual Alignment and Aggregation module that integrates Ref features that have been warped from matching maps obtained from two distinct sources. Given the challenges associated with training a precise patch matching for adapting the model to real-world input in a self-supervised manner, we leverage geometric matching maps to facilitate image matching between image pairs that exist a substantial FoV gap. Our experimental results demonstrate that our approach achieves state-of-the-art performance on both LR and real-world inputs.

Acknowledgments: This work was partly supported by JSPS KAKENHI Grant Number 20H05952 and 19H01110.

References

[1] 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. 1, 2

[2] Jianchun Chen, Lingjing Wang, Xiang Li, and Yi Fang. Arbicon-net: Arbitrary continuous geometric transformation

networks for image registration. *Advances in neural information processing systems*, 32, 2019. 2

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[4] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2, 5

[5] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5931–5940, 2022. 1, 2

[6] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018. 2

[7] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2

[8] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021. 2, 4, 5

[9] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12339–12348, 2019. 2

[10] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. *arXiv preprint arXiv:2203.14537*, 2022. 1, 2, 4, 6

[11] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[12] 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*, pages 768–783, 2018. 2

[13] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1034–1042. IEEE, 2019. 2

[14] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 346–363. Springer, 2020. 2

[15] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 2

[16] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018. 2

[17] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 2

[18] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–364, 2018. 2

[19] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2

[20] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1408–1423, 2019. 2

[21] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015. 2

[22] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[23] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020. 2

[24] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6258–6268, 2020. 2, 3

[25] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *arXiv preprint arXiv:2109.13912*, 2021. 6

[26] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10346–10356, 2021. 2

[27] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8708–8718, 2022. 2

[28] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2001–2010, 2021. 1, 2, 4, 6, 7, 8

[29] 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 *Proceedings of European Conference on Computer Vision*, pages 230–245, 2020. 1

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

[31] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of European Conference on Computer Vision*, 2018. 6, 7, 8

[32] 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

[33] 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 *Proceedings of British Machine Vision Conference*, 2017. 1

[34] 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*, pages 88–104, 2018. 1, 2