

DM³Net: Dual-Camera Super-Resolution via Domain Modulation and Multi-scale Matching

Cong Guan^{1,*}, Jiacheng Ying^{2,*}, Yuya Ieiri¹, Osamu Yoshie^{1,†}

¹Waseda University, Japan ²Zhejiang University, China

*Equal contribution †Corresponding author

guancong@fuji.waseda.jp yingjiacheng@zju.edu.cn ieyuharu@ruri.waseda.jp yoshie@waseda.jp

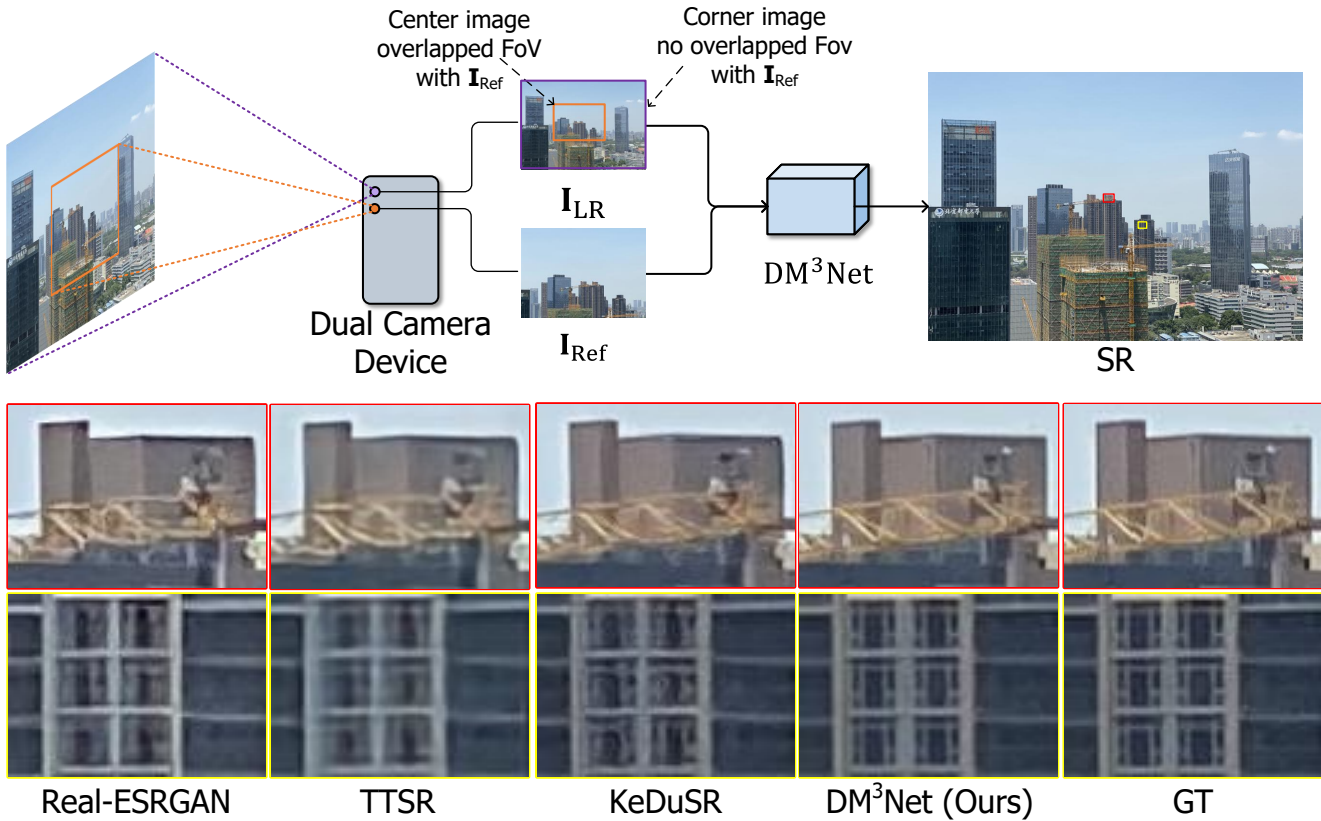


Figure 1. Modern multi-lens devices capture scenes at multiple focal lengths simultaneously. By fusing images from these diverse focal ranges, overall image quality and visual appeal can be substantially improved. The proposed method, DM³Net, achieves state-of-the-art performance with clear advantages in perceptual quality.

Abstract

Dual-camera super-resolution is highly practical for smart-phone photography that primarily super-resolve the wide-angle images using the telephoto image as a reference. In this paper, we propose DM³Net, a novel dual-camera super-resolution network based on **Domain Modulation** and

Multi-scale Matching. To bridge the domain gap between the high-resolution domain and the degraded domain, we learn two compressed global representations from image pairs corresponding to the two domains. To enable reliable transfer of high-frequency structural details from the reference image, we design a multi-scale matching module that conducts patch-level feature matching and retrieval across

multiple receptive fields to improve matching accuracy and robustness. Moreover, we also introduce Key Pruning to achieve a significant reduction in memory usage and inference time with little model performance sacrificed. Experimental results on three real-world datasets demonstrate that our DM³Net outperforms the state-of-the-art approaches.

1. Introduction

With the advancement of multi-camera device hardware (e.g., smartphones and action cameras), it has become possible to simultaneously capture images at different focal lengths and resolutions using both wide-angle and telephoto lenses on the same device. The wide-angle lens produces images with a larger FOV (Field of View) but lower resolution, while the telephoto lens captures a narrower FOV with higher resolution. The telephoto image, which contains high-frequency details, can serve as a reference for the super-resolution of wide-angle images. This task is a sub-problem in the Reference-based Super-Resolution (RefSR) field, and is referred to as Dual-Camera Super-Resolution [21].

Due to the content inconsistency between telephoto and wide-angle images, previous works have exploited patch matching mechanisms to transfer high-fidelity patches from the reference image to the target image. For instance, DCSR [21] performs feature matching between the reference and LR images to locate the most relevant reference patch for each LR patch, enabling detail fusion. Similarly, KeDuSR [30] matches the LR and LR-center images to locate the optimal patch positions, and uses the corresponding reference patches to reassemble a high-resolution feature. However, the matching strategies adopted by previous works are generally implemented at a coarse scale, which limits the receptive field and leads to a loss of fine details during reconstruction.

To resolve this, we propose a multi-scale matching mechanism. Specifically, we extract multi-scale features from the LR and LR-center images as the Query (**Q**) and Key (**K**) features. After applying feature patchification, we perform **Q-K** patch matching and reference feature re-assembly at each scale. The resulting high-resolution features from all scales are then fused to enhance the final reconstruction. Moreover, since many key patches represent highly similar regions and cause redundant matching operations, we introduce a Key Pruning strategy to suppress unnecessary computations. This significantly reduces inference time and memory usage with minimal degradation in performance.

In addition to the local matching strategy for detail refinement, we further exploit a global domain modulation mechanism. We learn two global domain-aware embeddings from the LR-GT image pair and the LR center-Ref

image pair to model the global domain gap between the high-resolution and degraded domains. Since the two embeddings are extracted from different image contents, we introduce a domain-aware loss to encourage their distribution similarity to filter out the content impact. These embeddings are then used to modulate the image reconstruction process, effectively improving the reconstruction accuracy.

Built upon the above multi-scale matching and domain modulation mechanisms, we propose a novel dual-camera super-resolution network, named DM³Net. Fig.1 presents example results generated by Real-ESRGAN [22], TTSR [28], KeDuSR [30], and our DM³Net. It can be observed that our method achieves noticeably better accuracy in reconstructing fine details compared to the others.

Our contributions can be summarized as follows:

- We propose DM³Net, a novel dual-camera super-resolution network based on domain modulation and multi-scale matching. Extensive experiments on three datasets validate that the proposed DM³Net outperforms the existing approaches in both center region and corner region. Moreover, the proposed DM³Net also has the best generalization ability in cross-dataset evaluation.
- To transfer high-fidelity details from reference image, we design a multiscale matching module that search the best-matching reference patches across different receptive fields for high-resolution feature reassembly and fusion. To improve the model efficiency, we introduce Key Pruning that eliminates redundant key patches during the matching process, significantly reducing both memory usage and inference time with little performance degradation.
- To globally bridge the domain gap between the high-resolution domain and the degraded domain, we introduce a domain modulation mechanism that learns two global domain-aware embeddings from the LR-GT pair and the LR center-Ref pair. Each of the embeddings is used to modulate the image reconstruction process, effectively improving the reconstruction accuracy.

2. Related Work

2.1. Reference-based Super-Resolution (RefSR)

RefSR methods enhance an LR image using an HR reference with similar content, where accurate alignment or matching is crucial. Early works [14, 18, 28, 29, 32, 34] rely on optical flow-based warping. SADC-SR [18] employs deformable convolution [35], while SRNTT [32] and C²-Matching [11] adopt patch-wise feature matching to better handle misalignment. Transformer-based methods [2, 15, 28] use multi-head attention for adaptive fusion. MASA-SR [15] and AMSA [25] further improve matching via coarse-to-fine and contrastive strategies. These approaches highlight the benefit of effective texture transfer

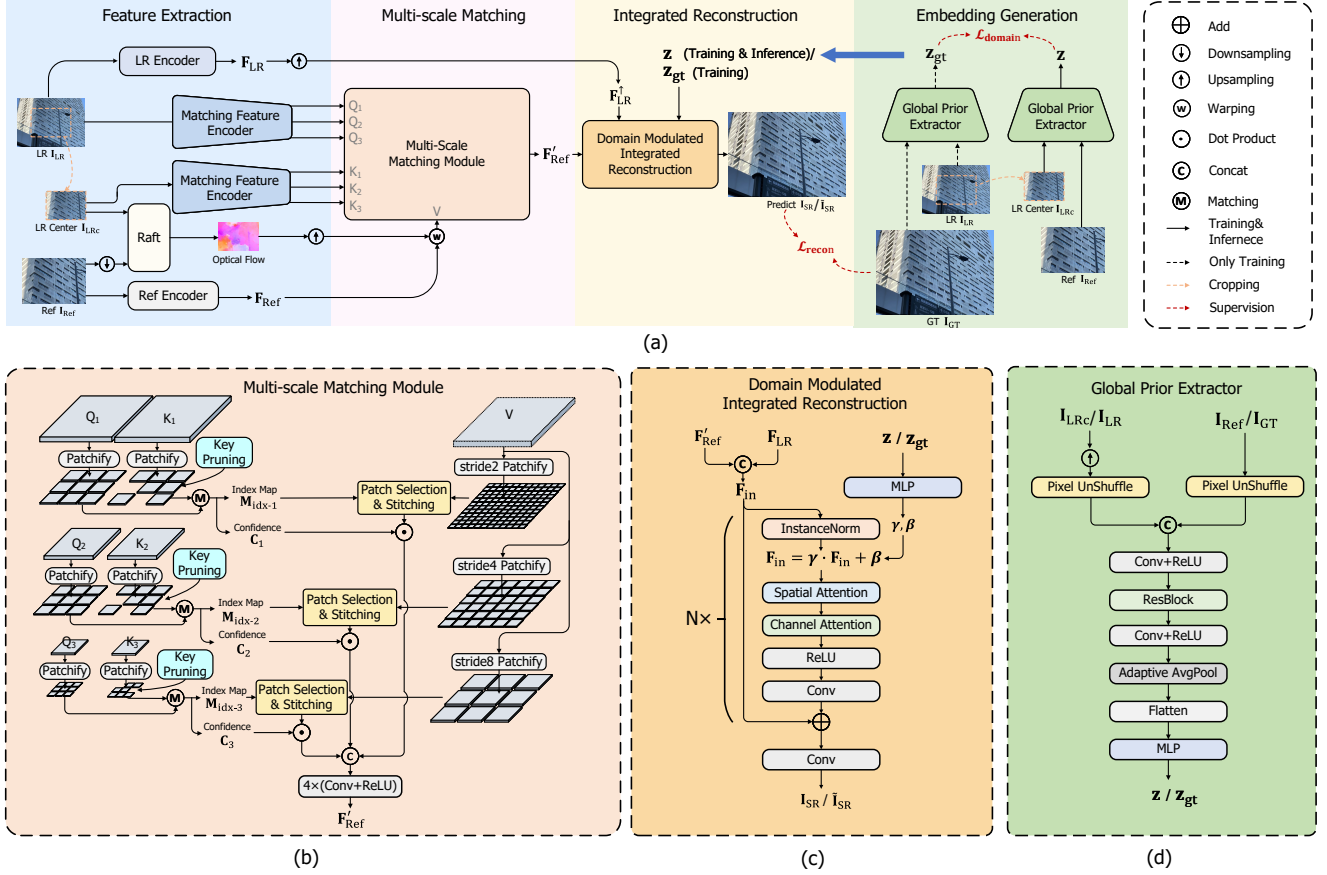


Figure 2. Schematics and detailed architectures of DM³Net. (a) The framework of the proposed DM³Net for dual camera super-resolution. The pipeline consists of four image encoders for feature extraction, the RAFT module for dense registration, the multi-scale matching (MSM) module for feature matching and reassembly, the global prior extractors (GPE) for domain-aware embeddings generation, and the domain modulated integrated reconstruction (DMIR) for resultant image prediction. (b) Detailed structure of multi-scale matching module. (c) Details of domain modulated integrated reconstruction. (d) Details of global prior extractor.

from aligned references.

2.2. Dual-Camera Super-Resolution

Dual-camera super-resolution is highly practical for mobile photography, where images are simultaneously captured using different lenses. In this scenario, the telephoto image serves as the reference to enhance the wide-angle LR image. Traditional approaches [4, 16, 17] often involve brightness and color correction along with global registration to mitigate differences between the images. Recently, deep learning methods address challenges like partial overlap, misalignment, and domain gaps in dual-camera SR. DCSR [21] pioneered deep models for real-world dual-camera SR, integrating telephoto details and alignment for spatial and color correction. Self-supervised approaches like SelfDZSR [33] and ZeDuSR [27] use telephoto images as supervision, eliminating the need for external HR labels. Further, works such as [24, 30] improve efficiency and ro-

bustness through optimized system and matching designs.

3. Method

3.1. Overview

Fig. 2 (a) illustrates the framework of the proposed DM³Net, which consists of four stages, *i.e.* feature extraction, multi-scale matching, embedding generation, and integrated reconstruction. Given a wide-angle image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$ and a telephoto reference image $I_{Ref} \in \mathbb{R}^{H \times W \times 3}$, we first crop the LR center region $I_{LRc} \in \mathbb{R}^{h \times w \times 3}$ that has the same FOV with I_{Ref} . The spatial dimensions are related with the scale factor d as $H = d \times h$ and $W = d \times w$. The target of DM³Net is to predict a HR image $I_{SR} \in \mathbb{R}^{dH \times dW \times 3}$ with richer structure details.

To achieve this, we extract features from I_{LR} , I_{LRc} , and I_{Ref} , separately. Optical flow between I_{LRc} and I_{Ref} is estimated to guide the feature alignment. Multi-scale matching is then performed between I_{LR} and I_{LRc} based on patch

similarity to search the best-matching reference patches for high-resolution feature reassembly. In parallel, two global domain-aware embeddings $\mathbf{z}$ and $\mathbf{z}_{\text{gt}}$ are extracted from the $\mathbf{I}_{\text{LRc}}\text{-}\mathbf{I}_{\text{Ref}}$ pair and $\mathbf{I}_{\text{LR}}\text{-}\mathbf{I}_{\text{GT}}$ pair to modulate the reconstruction process. These two embeddings are encouraged to follow a similar distribution through a domain-aware loss during training. Finally, the LR features $\mathbf{F}_{\text{LR}}^\uparrow$, the Ref feature $\mathbf{F}_{\text{Ref}}'$ and the domain-aware embeddings are fed into a reconstruction module to generate the output image $\mathbf{I}_{\text{SR}}$.

3.2. Feature Extraction

LR Encoder and Ref Encoder: The LR encoder and Ref encoder share a similar architecture, each consisting of several residual blocks [6] and channel attention modules [8]. These encoders are designed to extract the LR feature $\mathbf{F}_{\text{LR}} \in \mathbb{R}^{C_1 \times H \times W}$ and the Ref feature $\mathbf{F}_{\text{Ref}} \in \mathbb{R}^{C_1 \times H \times W}$ from $\mathbf{I}_{\text{LR}}$ and $\mathbf{I}_{\text{Ref}}$, respectively. The extracted LR feature $\mathbf{F}_{\text{LR}}$ is then upsampled to $\mathbf{F}_{\text{LR}}^\uparrow \in \mathbb{R}^{C_1 \times dH \times dW}$ to match the target resolution.

RAFT: To address the misalignment between $\mathbf{I}_{\text{LRc}}$ and $\mathbf{I}_{\text{Ref}}$, we employ a pretrained RAFT model (small version) [20] to estimate the optical flow $\mathbf{O} \in \mathbb{R}^{2 \times h \times w}$ at a coarse scale. The estimated optical flow is then upsampled to $H \times W$ and used to warp the Ref feature $\mathbf{F}_{\text{Ref}}$, resulting in the aligned feature $\mathbf{V} \in \mathbb{R}^{C_1 \times H \times W}$:

$$\mathbf{V} = \text{Warp}(\mathbf{F}_{\text{Ref}}, \mathbf{O}^\uparrow), \quad (1)$$

where $\uparrow$ denotes the upsampling operation. We note that the RAFT module is unfrozen during training to participate in the joint optimization.

Matching Feature Encoder: We use the pretrained VGG-19 [19] network as the matching feature encoder and take the three hidden layers as the multi-scale features. We extract the multi-scale $\mathbf{Q}$ and $\mathbf{K}$ features from $\mathbf{I}_{\text{LR}}$ and $\mathbf{I}_{\text{LRc}}$, respectively, using the weight-shared matching feature encoders, formulated as

$$\begin{aligned} \mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3 &= \text{MFE}(\mathbf{I}_{\text{LR}}), \\ \mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3 &= \text{MFE}(\mathbf{I}_{\text{LRc}}). \end{aligned} \quad (2)$$

where $\mathbf{Q}_1 \in \mathbb{R}^{C_2 \times H \times W}$, $\mathbf{Q}_2 \in \mathbb{R}^{C_2 \times \frac{H}{2} \times \frac{W}{2}}$, $\mathbf{Q}_3 \in \mathbb{R}^{C_2 \times \frac{H}{4} \times \frac{W}{4}}$, $\mathbf{K}_1 \in \mathbb{R}^{C_2 \times h \times w}$, $\mathbf{K}_2 \in \mathbb{R}^{C_2 \times \frac{h}{2} \times \frac{w}{2}}$, $\mathbf{K}_3 \in \mathbb{R}^{C_2 \times \frac{h}{4} \times \frac{w}{4}}$, and $\text{MFE}(\cdot)$ denote the matching feature encoder.

3.3. Multi-scale Matching

The feature of LR center region can be complemented by the aligned high-resolution Ref features. However, in the corner regions, it is difficult to directly obtain corresponding high-resolution features. To address this issue, previous approaches [21, 30] adopt block matching strategy to search similar feature patches in the center region and use them to enhance the feature patches in the corner region.

Table 1. The performance, inference time (IT), and memory usage (MU) before and after using Key Pruning. Tests are conducted on NVIDIA H20 on CameraFusion-Real dataset.

	PSNR $\uparrow$ / SSIM $\uparrow$	IT (s)	MU (GB)
Before Key Pruning	27.87 / 0.8436	31.08	49.41
After Key Pruning	27.75 / 0.8364	11.32	40.95

However, these methods typically perform single-scale feature matching at a coarse scale, which restricts the receptive field of the matching process. In this work, we propose a multi-scale matching module to further improve the accuracy and robustness of feature transfer from the Ref image. As illustrated in Fig. 2 (b), the multi-scale matching module takes the $\mathbf{V}$ feature, multi-scale $\mathbf{Q}$ and $\mathbf{K}$ features as the input. For the i -th scale, we patchify $\mathbf{V}$, $\mathbf{Q}_i$ and $\mathbf{K}_i$ to obtain patches of $\mathbf{V}$, $\mathbf{Q}$ and $\mathbf{K}$. And then, we compute cosine similarity between each $\mathbf{Q}$ patches and all $\mathbf{K}$ patches to select the best-matching $\mathbf{K}$ patch to form both a matching index map $\mathbf{M}_{\text{idx-}i}$ and a confidence map $\mathbf{C}_i$, as illustrated in Fig. 3 (a). Each element of $\mathbf{M}_{\text{idx-}i}$ and $\mathbf{C}_i$ can be computed as

$$M_{\text{idx-}i}^j = \arg \max_{1 \leq k \leq N} \text{sim}(\mathbf{Q}_i^j, \mathbf{K}_i^k), \quad (3)$$

$$C_i^j = \max_{1 \leq k \leq N} \text{sim}(\mathbf{Q}_i^j, \mathbf{K}_i^k) \quad (4)$$

where j denotes the element position, N denotes the number of $\mathbf{K}$ patches, $\text{sim}(\cdot)$ denotes the cosine similarity operator. The index map $\mathbf{M}_{\text{idx-}i}$ is then used to retrieve the corresponding $\mathbf{V}$ patches and reassemble a high-resolution feature, while the confidence map $\mathbf{C}_i$ is applied via element-wise multiplication to adaptively weight the stitched features. Finally, we concatenate the three weighted feature maps and feed them to a fusion head that consists of four convolution and ReLU[1] layers to produce the output feature $\mathbf{F}_{\text{Ref}}' \in \mathbb{R}^{C_1 \times dH \times dW}$.

Fig 4 presents an example image generated by KeDuSR [30], which employs single-scale matching, and our method. It is observed our method outperforms KeDuSR in terms of structural consistency and fine details. This improvement is primarily attributed to the enhanced reliability of feature transfer enabled by our multi-scale matching mechanism.

However, we notice that the above multi-scale matching mechanism between the $\mathbf{Q}$ and $\mathbf{K}$ features incurs significant memory consumption and increases inference time. Considering that many $\mathbf{K}$ patches are highly similar to some others, a large portion of the similarity computations between $\mathbf{Q}$ and $\mathbf{K}$ patches are redundant and unnecessary. Motivated by this observation, we propose Key Pruning to improve the efficiency of the model, as illustrated in Fig. 3 (b). We uniformly sample a subset of $\mathbf{K}$ patches, compute the similarity between the sampled patches and all $\mathbf{K}$ patches,

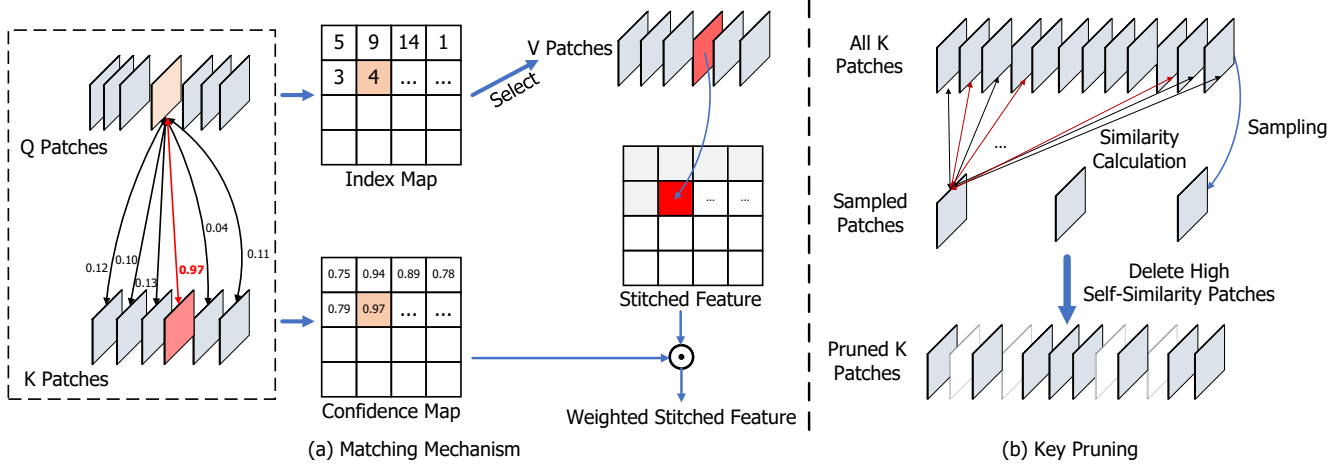


Figure 3. The flow of (a) the Matching Mechanism and (b) the Key Pruning.



Figure 4. Reconstructed images produced by the single-scale matching approach, KeDuSR [30], and our proposed method. Our method achieves better preservation of texture details.

and remove those patches whose similarity exceeds a certain threshold. The remaining patches serve as the pruned K patches for subsequent matching. In this work, we set the sampling interval as 16 in both the horizontal and vertical directions, and set the threshold as 0.7. Through Key Pruning, we achieve a substantial reduction in resource consumption by sacrificing little model performance. Specifically, as listed in table 1, on the CameraFusion-Real dataset,

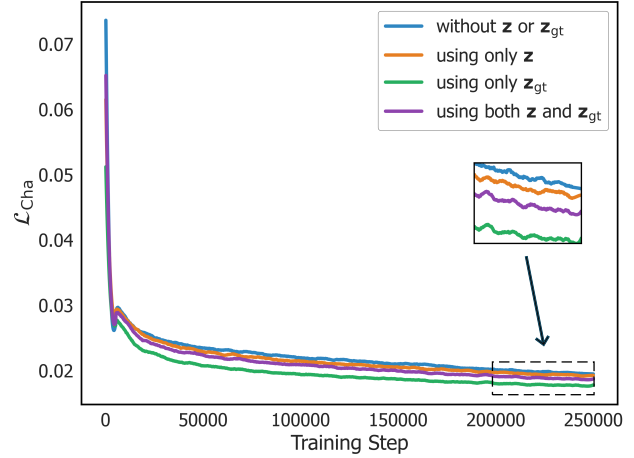


Figure 5. Visualization of the Charbonnier loss [3] curves during training under four experiment settings

memory usage was reduced from 49.41 GB to 40.95 GB, and inference time decreased from 31.08s to 11.32s, with only a slight performance drop of 0.12 dB in PSNR[7] and 0.072 in SSIM[23].

3.4. Embedding Generation

Motivated by previous image restoration approaches [9, 26], we also learn a compressed global representation that captures the high-fidelity prior to bridge the domain gap between the high-resolution domain and the degraded domain. To this end, we propose *Global Prior Extractor* (GPE) that extracts a domain-aware embedding from two images corresponding to the two domains. The structure of GPE is illustrated in Fig. 2 (d). In particular, we consider two different input pairs: $\{\mathbf{I}_{LR}, \mathbf{I}_{GT}\}$ and $\{\mathbf{I}_{LRc}, \mathbf{I}_{Ref}\}$. Both pairs share two key characteristics: (1) the two images have the same scene region, and (2) one in the high-resolution do-

main and the other in the degraded domain. We employ two GPEs to extract domain-aware embeddings $z \in \mathbb{R}^{1 \times 1024}$ and $z_{gt} \in \mathbb{R}^{1 \times 1024}$ using the pair $\{\mathbf{I}_{LRc}, \mathbf{I}_{Ref}\}$ and pair $\{\mathbf{I}_{LR}, \mathbf{I}_{GT}\}$, respectively, formulated as

$$z = \text{GPE}_1(\mathbf{I}_{LRc}, \mathbf{I}_{Ref}), \quad z_{gt} = \text{GPE}_2(\mathbf{I}_{LR}, \mathbf{I}_{GT}). \quad (5)$$

Each of the embeddings is employed in the Integrated Reconstruction stage for image reconstruction.

Since z_{gt} is extracted from the GT image, the Charbonnier loss, one of the image reconstruction loss, of the model using z_{gt} during training is significantly lower than that using z (see the orange and blue curves in Fig. 5). However, z_{gt} is unavailable during inference because the GT image is not accessible. To mitigate this gap, we employ the domain-aware loss $\mathcal{L}_{\text{domain}}$ to enforce z to have similar distribution to z_{gt} , which also helps reduce the image reconstruction loss (see the green curve in Fig. 5). In this way, during inference, we can use z as a substitute for z_{gt} for image reconstruction.

3.5. Integrated Reconstruction

In this stage, the reassembled ref feature $\mathbf{F}'_{\text{Ref}}$, the LR feature $\mathbf{F}_{LR}^\uparrow$, and the domain-aware embedding z/z_{gt} are fed into the Domain Modulated Integrated Reconstruction (DMIR) module to predict the super-resolved image $\mathbf{I}_{SR}/\tilde{\mathbf{I}}_{SR}$. The architecture of DMIR module is illustrated in Fig. 2 (c), where the domain-aware embedding z/z_{gt} is processed by an MLP[5] to generate affine parameters γ and β for feature modulation [10].

We note that during the training phase, $\mathbf{I}_{SR}$ denotes the output using z as the input, while $\tilde{\mathbf{I}}_{SR}$ denotes the output using z_{gt} . Both outputs are supervised to encourage z and z_{gt} to effectively bridge the high-resolution and degraded domain gap. During inference, only z is required as input to generate the predicted image $\mathbf{I}_{SR}$.

3.6. Loss Functions

In the training phase, we employ reconstruction loss $\mathcal{L}_{\text{rec}}$ and domain-aware loss $\mathcal{L}_{\text{domain}}$ to guide the learning process. The reconstruction loss consists of the Charbonnier loss $\mathcal{L}_{\text{cha}}$ and the perceptual loss $\mathcal{L}_{\text{per}}$, while the domain-aware loss $\mathcal{L}_{\text{domain}}$ is used to encourage the two domain-aware embeddings to have similar distributions.

Charbonnier Loss. We use the Charbonnier loss [3], a differentiable variant of the L_1 loss, as the primary reconstruction loss for both predicted images $\mathbf{I}_{SR}$ and $\tilde{\mathbf{I}}_{SR}$. The loss is defined as:

$$\mathcal{L}_{\text{cha}} = \sqrt{(\mathbf{I}_{SR} - \mathbf{I}_{GT})^2 + \epsilon^2} + \sqrt{(\tilde{\mathbf{I}}_{SR} - \mathbf{I}_{GT})^2 + \epsilon^2}, \quad (6)$$

where ϵ is a small constant (set to 10^{-6} in our experiments) added for numerical stability.

Perceptual Loss. To further enhance the image visual fidelity, we introduce the perceptual loss [12] that compares deep features extracted from a pretrained VGG network [19]. The perceptual loss is defined as:

$$\mathcal{L}_{\text{percep}} = \sum_l \|\phi_l(\mathbf{I}_{SR}) - \phi_l(\mathbf{I}_{GT})\|_2^2 + \sum_l \|\phi_l(\tilde{\mathbf{I}}_{SR}) - \phi_l(\mathbf{I}_{GT})\|_2^2, \quad (7)$$

where $\phi_l(\cdot)$ denotes the activation of the l -th layer of the VGG network.

Domain-Aware Loss. We employ the domain-aware loss to encourage the domain-aware embeddings z and z_{gt} to have similar distributions, computed as

$$\mathcal{L}_{\text{domain}} = \|z - z_{gt}\|_1. \quad (8)$$

Total Loss. The final training objective is formulated as a weighted combination of all loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{domain}} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{per}} + \lambda_3 \mathcal{L}_{\text{domain}}, \quad (9)$$

where λ_{rec} , λ_{percep} , and λ_{embed} are balancing parameters. We empirically set $\lambda_1 = 1$, $\lambda_2 = 0.01$, and $\lambda_3 = 1000$ in this work.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct our experiments on 3 public datasets: **DuSR-Real**, **RealMCVSR-Real**, and **CameraFusion-Real**. Details of the datasets are in Supplementary Materials.

Comparing Methods. We compare our method with single image SR approaches, reference-based SR approaches, and dual-camera SR approaches. The single image SR approaches include SwinIR [13] and Real-ESRGAN [22]. The reference-based SR includes MASA-SR [15] and TTSR [28]. The dual-camera SR approaches include DCSR [21], SelfDZSR [33], and KeDuSR [30]. For a fair comparison, we retrain all the aforementioned approaches except KeDuSR [30] using the same training set as our method.

4.2. Small Version of DM³Net

To meet the memory constraints of edge deployment, we further propose a lightweight variant of DM³Net. The design follows the principle of reducing model complexity while maintaining performance. The resulting model, DM³Net-small, contains only **40.4%** of the parameters of the original version. Despite this substantial reduction, evaluations on three benchmark datasets demonstrate that the performance drop is marginal, and the model still achieves **state-of-the-art** results. Please refer to the supplementary materials for more details.

Table 2. Quantitative comparisons on DuSR-Real dataset. The best metrics are in bold.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Center PSNR/SSIM	Corner PSNR/SSIM	Inference Time (s)	Params (M)
SwinIR	26.02	0.8580	0.226	26.22 / 0.8592	25.99 / 0.8576	4.609	11.75
Real-ESRGAN	24.19	0.8339	0.188	24.39 / 0.8372	24.17 / 0.8328	0.113	16.70
MASA-SR	26.17	0.8528	0.233	26.69 / 0.8554	26.05 / 0.8519	7.067	6.25
TTSR	25.59	0.8544	0.234	26.14 / 0.8612	25.46 / 0.8521	6.013	4.02
DCSR	26.06	0.8576	0.190	28.33 / 0.8975	25.54 / 0.8440	1.172	3.19
SelfDZSR	25.62	0.8351	0.198	26.23 / 0.8387	25.47 / 0.8339	0.793	0.52
KeDuSR	27.18	0.8752	0.164	29.06 / 0.9219	26.77 / 0.8593	0.836	5.63
DM³Net	27.74	0.8879	0.159	29.67 / 0.9291	27.34 / 0.8738	1.404	18.64
DM³Net-small	27.65	0.8875	0.161	29.37 / 0.9302	27.28 / 0.8729	1.387	7.53

Table 3. Quantitative comparisons on RealMCVSR-Real and CameraFusion-Real datasets. The best metrics are in bold.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Center PSNR/SSIM	Corner PSNR/SSIM
RealMCVSR-Real					
SwinIR	25.78	0.7992	0.304	25.57 / 0.7896	25.92 / 0.8025
Real-ESRGAN	24.00	0.7704	0.251	23.79 / 0.7640	24.15 / 0.7725
MASA-SR	25.81	0.7911	0.315	25.73 / 0.7821	25.90 / 0.7941
TTSR	24.83	0.7889	0.311	25.63 / 0.7990	24.63 / 0.7855
DCSR	25.79	0.8024	0.275	27.14 / 0.8524	25.47 / 0.7853
SelfDZSR	25.15	0.7803	0.276	25.32 / 0.7748	25.17 / 0.7821
KeDuSR	26.42	0.8184	0.226	28.51 / 0.9090	25.95 / 0.7875
DM³Net	27.07	0.8394	0.222	29.27 / 0.9197	26.58 / 0.8120
DM³Net-small	26.90	0.8356	0.229	29.10 / 0.9145	26.40 / 0.8087
CameraFusion-Real					
SwinIR	25.32	0.7926	0.347	25.90 / 0.8000	25.19 / 0.7901
Real-ESRGAN	24.37	0.7643	0.287	24.79 / 0.7690	24.30 / 0.7626
MASA-SR	25.92	0.7886	0.343	26.75 / 0.8004	25.74 / 0.7846
TTSR	25.51	0.7945	0.376	25.42 / 0.7890	25.61 / 0.7964
DCSR	24.96	0.7569	0.316	26.15 / 0.7898	24.67 / 0.7458
SelfDZSR	25.78	0.7833	0.280	26.76 / 0.7950	25.55 / 0.7794
KeDuSR	27.00	0.7931	0.274	29.77 / 0.8418	26.43 / 0.7768
DM³Net	27.75	0.8364	0.131	32.22 / 0.8894	27.08 / 0.8186
DM³Net-small	27.55	0.8387	0.280	31.48 / 0.9090	26.80 / 0.8147

4.3. Quantitative Comparison

Table 2 and Table 3 present the quantitative results on 3 benchmark datasets. Full-Image denotes evaluation over the entire image, Center-Image corresponds to the overlapping FOV region, and Corner-Image refers to the non-overlapping regions. Across all datasets, our DM³Net consistently achieves the best Full-Image performance in PSNR, SSIM, and LPIPS [31], demonstrating its superior reconstruction fidelity. Moreover, the performance advantage of our method remains evident in both Center-Image and Corner-Image regions, especially on the CameraFusion-Real dataset, highlighting its strong restoration capability and robustness to FoV variation.

4.4. Qualitative Comparison

Figure 6 shows qualitative comparisons on three benchmark datasets. SwinIR, TTSR, MASA-SR, and SelfZeDuSR often produce blurry textures and structural distortions (e.g., red boxes in CameraFusion-Real, yellow in DUSR-Real,

red in RealMCVSR-Real). TTSR also exhibits color distortion (e.g., DUSR-Real). While Real-ESRGAN, DCSR, and KeDuSR generate sharper details, they introduce artifacts around characters (see red/yellow boxes), and KeDuSR further fails to preserve window and block shapes. In contrast, DM³Net reconstructs high-frequency textures and character details faithfully without color distortion. Additional comparisons are provided in the Supplementary Materials.

4.5. Ablation Study

Ablation studies are conducted on the CameraFusion-Real dataset to assess the impact of the proposed multi-scale matching and domain modulation.

Multi-Scale Matching: To assess the impact of the multi-scale matching, we evaluated seven different combinations of matching at 1/4 scale, 1/2 scale, and full scale. As listed in Table 4, adopting higher-scale matching generally leads to consistently better reconstruction accuracy. Moreover, combining multiple scales brings further improvements, be-

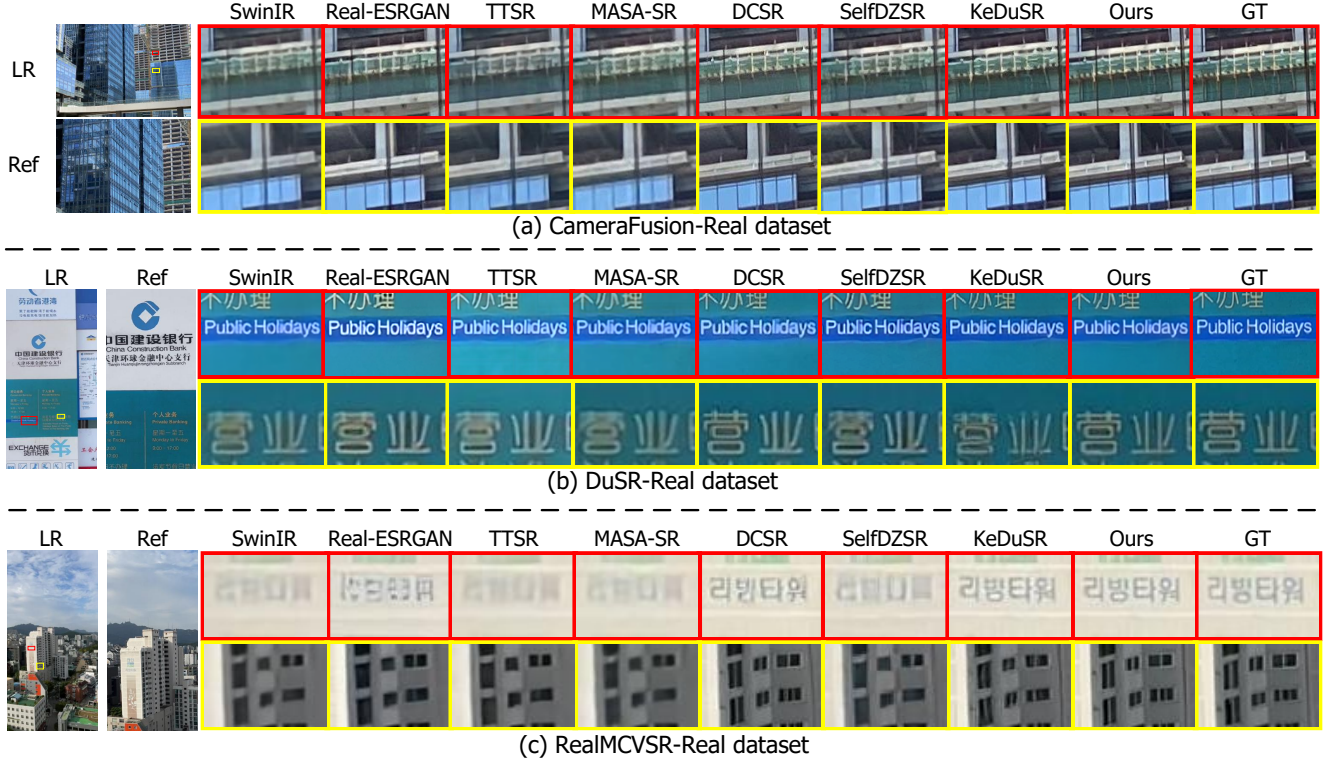


Figure 6. Qualitative comparison results on 3 datasets. Zoom in for best comparison.

Table 4. Ablation Study on Multi-scale Matching. 1/4, 1/2, and 1 denote matching at respective scales.

1/4	1/2	1	PSNR $\uparrow$	SSIM $\uparrow$
✓			26.55	0.8112
	✓		27.17	0.8250
		✓	27.30	0.8327
✓	✓		27.41	0.8331
✓		✓	27.55	0.8335
	✓	✓	27.54	0.8365
✓	✓	✓	27.75	0.8364

Table 5. Ablation Study on Domain Modulation.

$\mathbf{z}$	$\mathbf{z}_{gt}$	PSNR $\uparrow$	SSIM $\uparrow$
55	55	27.37	0.8337
✓	55	27.50	0.8360
✓	✓	27.75	0.8364

cause the receptive fields associated with different scales are highly complementary and effectively mitigate the inherent limitations of each individual scale.

Domain Modulation: To validate the effectiveness of the domain-aware embeddings $\mathbf{z}$ and $\mathbf{z}_{gt}$, we trained three versions of models: (1) a baseline model without $\mathbf{z}$ and $\mathbf{z}_{gt}$, (2) a model incorporating only $\mathbf{z}$ without supervision from

$\mathbf{z}_{gt}$, and (3) the full model. As listed in Table 5, incorporating $\mathbf{z}$ alone already improves the reconstruction metrics, demonstrating that a global domain-aware embedding benefits high-resolution image reconstruction. Furthermore, the full model achieves the best performance, indicating that supervision from $\mathbf{z}_{gt}$ further enhances reconstruction quality, highlighting the importance of explicitly modeling the domain gap for robust dual-camera super-resolution.

5. Conclusions

In this work, we propose DM³Net, a novel framework for dual-camera super-resolution. DM³Net utilizes multi-scale matching to transfer high-quality structure details from the reference image, and leverages domain modulation to bridge the HR and the degraded domain gap. By introducing a key pruning strategy, DM³Net also enhances computational efficiency with slight performance drop. Extensive experiments on three real-world datasets demonstrate that our method achieves SOTA performance across multiple visual quality metrics. By explicitly modeling domain and FOV discrepancies, DM³Net achieves more accurate detail reconstruction and delivers strong generalization and robustness across diverse real-world dual-camera super-resolution scenarios, making it a promising solution for practical deployment.

References

- [1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 4
- [2] Jiezhong Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *Proceedings of the European Conference on Computer Vision*, 2022. 2
- [3] P. Charbonnier et al. Two-dimensional signal and image processing. *IEEE Transactions on Image Processing*, 3(2):68–72, 1994. 5, 6
- [4] Xiaojie Guo, Yifan Li, Zhe Yu, and Jiaya Liu. Low-light color imaging via dual camera acquisition. In *Asian Conference on Computer Vision (ACCV)*, pages 123–138, 2020. 3
- [5] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. 6
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [7] A. Hore and D. Ziou. Image quality metrics: Psnr vs. ssim—putting structural similarity on a sound mathematical foundation. In *Proceedings of the 2010 International Conference on Pattern Recognition (ICPR)*, pages 2366–2369, 2010. 5
- [8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 4
- [9] Tao Hu, Qingsen Yan, Yuankai Qi, and Yanning Zhang. Generating content for hdr dehazing from frequency view. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25732–25741, 2024. 5
- [10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 6
- [11] Yuming Jiang, Kelvin C.K. Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c^2 -matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2
- [12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *Proceedings of the European Conference on Computer Vision*, pages 694–711, 2016. 6
- [13] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 6
- [14] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2022. 2
- [15] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6
- [16] Joo Park, Sung Kim, and Hyun Kang. Brightness and color correction for dual camera image registration. In *Proceedings of IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4, 2016. 3
- [17] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2564–2571, 2011. 3
- [18] 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*, 2020. 2
- [19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 4, 6
- [20] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, 2020. 4
- [21] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2001–2010, 2021. 2, 3, 4, 6
- [22] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops*. 2, 6
- [23] 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*, 13(4): 600–612, 2004. 5
- [24] Xiaotong Wu, Wei-Sheng Lai, Yichang Shih, Charles Herrmann, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Efficient hybrid zoom using camera fusion on mobile phones. *ACM Trans. Graph.*, 42(4):1–15, 2023. 3
- [25] Bin Xia, Yapeng Tian, Yucheng Hang, Wenming Yang, Qingmin Liao, and Jie Zhou. Coarse-to-fine embedded patchmatch and multi-scale dynamic aggregation for reference-based super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2768–2776, 2022. 2
- [26] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. *Proceedings of IEEE International Conference on Computer Vision*, 2023. 5
- [27] Ruikang Xu, Mingde Yao, and Zhiwei Xiong. Zeroshot dual-lens superresolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3
- [28] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bainig Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 6

- ence on Computer Vision and Pattern Recognition, pages 5791–5800, 2020. [2](#), [6](#)
- [29] 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. [2](#)
- [30] Huanjing Yue, Zifan Cui, Kun Li, and Jingyu Yang. Kedusr: real-world dual-lens super-resolution via kernel-free matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6881–6889, 2024. [2](#), [3](#), [4](#), [5](#), [6](#)
- [31] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. [7](#)
- [32] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. [2](#)
- [33] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, Yunjin Chen, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual zoomed observations. In *Proceedings of the European Conference on Computer Vision*, 2022. [3](#), [6](#)
- [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. [2](#)
- [35] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. [2](#)