Zero-Shot Dual-Lens Super-Resolution

Ruikang Xu* Mingde Yao* Zhiwei Xiong†
University of Science and Technology of China
{xurk, mdyao}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn

Abstract

The asymmetric dual-lens configuration is commonly available on mobile devices nowadays, which naturally stores a pair of wide-angle and telephoto images of the same scene to support realistic super-resolution (SR). Even on the same device, however, the degradation for modeling realistic SR is image-specific due to the unknown acquisition process (e.g., tiny camera motion). In this paper, we propose a zero-shot solution for dual-lens SR (ZeDuSR), where only the dual-lens pair at test time is used to learn an image-specific SR model. As such, ZeDuSR adapts itself to the current scene without using external training data, and thus gets rid of generalization difficulty. However, there are two major challenges to achieving this goal: 1) dual-lens alignment while keeping the realistic degradation, and 2) effective usage of highly limited training data. To overcome these two challenges, we propose a degradation-invariant alignment method and a degradation-aware training strategy to fully exploit the information within a single dual-lens pair. Extensive experiments validate the superiority of ZeDuSR over existing solutions on both synthesized and real-world dual-lens datasets. The implementation code is available at https://github.com/XrKang/ZeDuSR.

1. Introduction

Mobile devices such as smartphones are generally equipped with an asymmetric camera system consisting of multiple lenses with different focal lengths. As a common configuration, with a wide-angle lens and a telephoto lens, one can capture the same scene with different field-of-views (FoVs). Within the overlapped FoV, the wide-angle and telephoto images naturally store low-resolution (LR) and high-resolution (HR) counterparts for learning a realistic super-resolution (SR) model. This provides a feasible way to obtain an HR image with a large FoV at the same time, which is beyond the capability of the original device.

There are a few pioneer works along this direction, which are referred to as dual-lens/camera/zoomed SR interchangeably. Beyond traditional methods that simply transfer the telephoto content to the wide-angle view in the overlapped FoV, recent deep-learning-based methods enable resolution enhancement of the full wide-angle image. Specifically, Wang *et al.* [43] utilize the overlapped FoV of dual-lens image pairs to learn a reference-based SR model, where the wide-angle view is super-resolved by using the telephoto view as a reference. However, the external training data are synthesized with the predefined bicubic degradation, which leads to generalization difficulty when the trained SR model is applied to real devices. To narrow the domain gap between the training and inference stages, they further adopt a self-supervised adaptation strategy to fine-tune the pretrained model on real devices. On the other hand, Zhang *et al.* [53] adopt self-supervised learning to train an SR model directly on a dual-lens dataset to avoid the predefined degradation, with the assumption of consistent degradation on the same device.

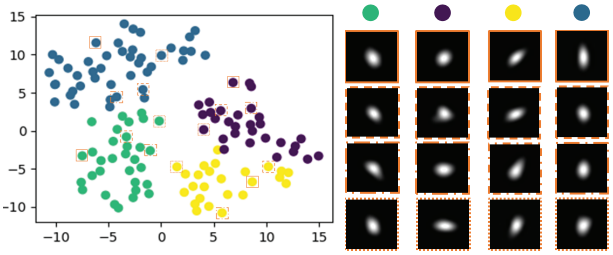


Figure 1. Degradation kernel (estimated by KernelGAN [1]) clustering on wide-angle images from iPhone11 [43].

In practice, however, the degradation kernel for modeling realistic SR is influenced by not only the camera optics but also tiny camera motion during the acquisition process on mobile devices [1, 6, 36], resulting in the image-specific degradation for each dual-lens pair, even on the same device. Figure 1 gives an exemplar analysis, where we estimate the degradation kernels of 146 wide-angle images captured by the dual-lens device iPhone11 [43] and cluster them using t-SNE [41]. Although these images are captured by the same device, they exhibit notably different degradation kernels due to the unknown acquisition process. This limits the performances of previous dual-lens SR methods for practical applications, since they assume that the realis-

*Equal contribution. †Corresponding author.

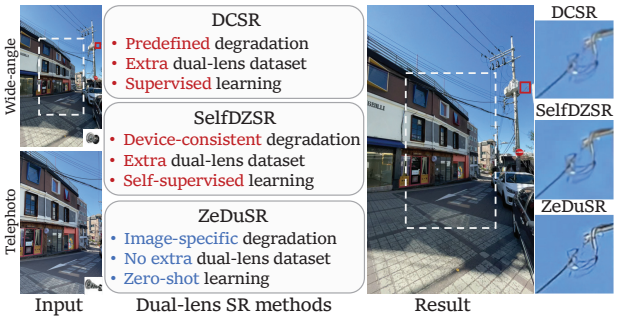


Figure 2. Overview of dual-lens SR methods and visual comparison of 2× SR on the real-world data.

tic degradation is at least device-consistent.

In this paper, we propose a zero-shot learning solution for realistic SR on dual-lens devices, termed as ZeDuSR, which learns an image-specific SR model solely from the dual-lens pair at test time. As such, ZeDuSR adapts itself to the current scene under the unknown acquisition process and gets rid of the generalization difficulty when using external training data. Figure 2 summarizes the main differences of ZeDuSR from existing dual-lens SR methods with a real-world reconstruction example. ZeDuSR gives a visually improved result compared with its competitors, thanks to the image-specific degradation assumption.

There are two major challenges to achieving the success of ZeDuSR. First, as also noticed in previous works [43,53], it is non-trivial to exploit the information of the dual-lens image pair, due to the spatial misalignment caused by the physical offset between the two lenses. Moreover, we find that the alignment process for generating the training data will introduce additional frequency information, which inevitably changes the realistic degradation. To overcome this challenge, we propose to constrain the alignment process in spatial, frequency, and feature domains simultaneously to keep the degradation. Specifically, we propose a degradation-invariant alignment method by leveraging adversarial and contrastive learning. The other challenge is how to effectively use the highly limited data for learning an image-specific SR model. To this end, we design a degradation-aware training strategy to fully exploit the information within a single dual-lens pair. That is, we calculate the probability of each location for patch cropping according to the degradation similarity between the images before and after alignment. As such, samples that keep the degradation are assigned a higher probability to be selected during training, while the contribution of degradation-variant samples is reduced.

We evaluate the performance of ZeDuSR on both synthesized and real-world dual-lens datasets. For quantitative evaluation with the HR ground-truth, we adopt a stereo dataset (Middlebury2021 [35]) and a light field dataset (HCI_new [12], with two views selected) to simulate dual-lens image pairs with different baselines between the two lenses, by applying image-specific degradation on one of the two views. Besides, we also perform the qualitative evaluation on two real-world datasets, *i.e.*, CameraFusion [43] captured by iPhone11 and RealMCVSR [18] captured by iPhone12, where no additional degradation is introduced beyond the realistic one between the two views. Extensive experiments on both synthesized and real-world datasets demonstrate the superiority of our ZeDuSR over existing solutions including single-image SR, reference-based SR, and dual-lens SR.

Contributions of this paper are summarized as follows:

• We propose a zero-shot learning solution for realistic SR on dual-lens devices, which assumes image-specific degradation and adapts itself to the current scene under the unknown acquisition process.

• We propose a degradation-invariant alignment method by leveraging adversarial and contrastive learning to constrain the alignment process in spatial, frequency, and feature domains simultaneously.

• We design a degradation-aware training strategy to effectively exploit the information within the highly limited training data, *i.e.*, a single dual-lens pair at test time.

• We conduct extensive experiments on both synthesized and real-world dual-lens datasets to validate the superiority of our zero-shot solution.

2. Related Work

2.1. Realistic Image SR

Based on the predefined degradation assumption, single image SR has seen significant advances in terms of both reconstruction accuracy [3, 9, 21, 30, 32] and perceptual quality [11,23] in the deep learning era, thanks to a large amount of simulated external training data. However, the difference between the predefined degradation and the realistic one brings great obstacles to applying the deep SR models trained with simulated data on real devices. Recently, much research attention has been drawn to the task of realistic SR [1, 7] (also known as blind SR). That is, the degradation kernel of the input LR image at test time is unknown.

Existing realistic SR methods characterize the realistic degradation from different perspectives, which can be roughly divided into four classes: (a) capture a large amount of well-aligned LR-HR image pairs and implicitly learn an SR model in a data-driven way [2, 4, 51], (b) estimate the degradation kernel from the input LR image itself with the self-similarity prior [1, 36, 38], (c) collect unpaired datasets in LR-HR spaces and learn the degradation in a circular fashion [48], and (d) simulate LR-HR image pairs using a predefined kernel pool consisting of various degradation kernels [10, 49]. Despite the corresponding merits, these methods suffer from respective compromises: (a) large effort of data collection for diverse devices, (b) unstable kernel estimation when the input image lacks self-similarity,

(c) and (d) difficulty to handle outliers beyond training data distribution. These drawbacks limit the performances of the above SR methods when the scene content and degradation kernel are unpredictable on real devices.

2.2. Dual-lens SR

Dual-lens SR aims to super-resolve wide-angle images with the assistance of telephoto images on the asymmetric camera system [5, 43], which generates HR and large FoV results beyond the capability of the original device. Traditional methods either adopt brightness and color correction under the assumption of geometric alignment between the two lenses [24, 33] or regard this task as similar to image registration [29] and simulate the HR result from the telephoto image through affine transformation. However, these methods fail to super-resolve the full wide-angle image and suffer from limited performance when applied to real data.

Recently, there emerge deep-learning-based methods for the task of reference-based SR, which aim to exploit the HR texture information from an additional reference image similar to the LR input to guide the SR process [13, 16, 25, 45–47, 54–57]. Subsequently, from this perspective, Wang *et al.* [43] take the telephoto image as the reference to training an SR model on a dual-lens dataset with the predefined bicubic degradation. To avoid the predefined degradation, Zhang *et al.* [53] train a self-supervised SR model directly from external dual-lens pairs with the assumption of device-consistent degradation. However, previous methods ignore that the realistic degradation is specific to each capture due to tiny camera motion in the acquisition process of mobile devices, which leads to generalization difficulty in practice.

3. Analysis of Realistic Dual-lens SR

3.1. Problem Formulation

We consider two images captured by a dual-lens device: the telephoto image X with a small FoV and the wide-angle image Y with a large FoV, as shown in Figure 4. Generally, within the overlapped FoV (denoted by the white dotted rectangle), X corresponds to a central area in Y (denoted as Y^*) but is with a much higher resolution. The decrease of resolution due to the enlarged FoV, which is termed Resolution-FoV (RV) degradation $D_{RV}(\cdot)$ in [4], can be characterized by the LR-HR image pair (Y^*, X).

In this paper, we argue that, the realistic RV degradation is not only dependent on the camera optics, but also influenced by tiny camera motion. Therefore, this degradation is image-specific and our goal is to obtain a parametric SR function $S_\Theta(\cdot)$ that reverses $D_{RV}(\cdot)$ for realistic SR, where Θ denotes the parameters of $S(\cdot)$. To achieve this goal, we optimize $S(\cdot)$ with the paired data (Y^*, X) as

$$\min_{\Theta} \mathcal{L}(X, S_\Theta(Y^*)), \tag{1}$$

where $\mathcal{L}(\cdot,\cdot)$ denotes a certain loss function. In this way, the implicit modeling of the image-specific degradation $D_{RV}(\cdot)$ is jointly achieved with the optimization of $S_\Theta(\cdot)$, since the training data is subjected to $Y^* = D_{RV}(X)$[1].

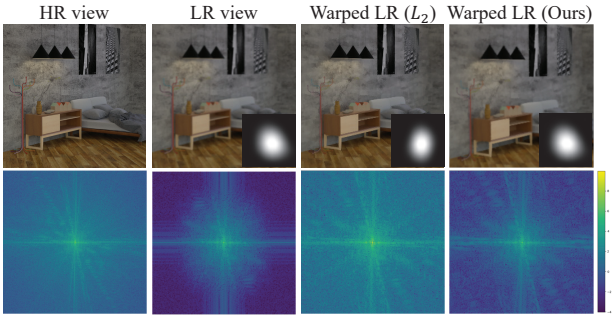


Figure 3. Analysis on the alignment process. The warped LR view using L_2 supervision introduces additional frequency information, resulting in degradation variance (revealed by estimated kernels). Our method adequately keeps the degradation and largely prevents the introduction of additional frequency information. Please see more analysis on frequency change in the supplement.

3.2. Challenge

As analyzed above, the LR-HR image pair (Y^*, X) from each dual-lens pair characterizes the image-specific degradation $D_{RV}(\cdot)$ for the current capture. Therefore, we propose to learn an image-specific SR model $S_\Theta(\cdot)$ by leveraging a single dual-lens pair at test time, which serves as a zero-shot learning solution. This solution addresses the generalization issue caused by the variation of realistic degradation from one capture to another, as encountered when using external training data.

Despite the clear advantage of zero-shot dual-lens SR, there are two major challenges to achieving this goal. First, the physical offset between the two lenses will lead to complicated misalignment between Y^* and X, especially for the scene with subjects at different depths. This can be addressed with advanced deep-learning-based alignment methods [14, 40], however, the alignment process will introduce additional frequency information and change the degradation $D_{RV}(\cdot)$ after alignment, as shown in Figure 3.

The other challenge is how to effectively use the highly limited data from a single dual-lens pair to learn the image-specific SR model $S_\Theta(\cdot)$. Even in a well-aligned image pair, there still exist a few regions with variant degradation. These noisy regions in such limited training data will lead to optimization difficulty and incorrect degradation modeling.

4. Zero-Shot Dual-Lens SR

An overview of our proposed zero-shot learning solution for dual-lens SR (ZeDuSR) is shown in Figure 4(a). During the training stage, we first downsample the HR telephoto image X (by the bilinear operator) to obtain the same resolution as its LR wide-angle counterpart Y^*, denoted as $X^{\downarrow}$.

[1] Strictly speaking, this excludes the dual-lens misalignment.

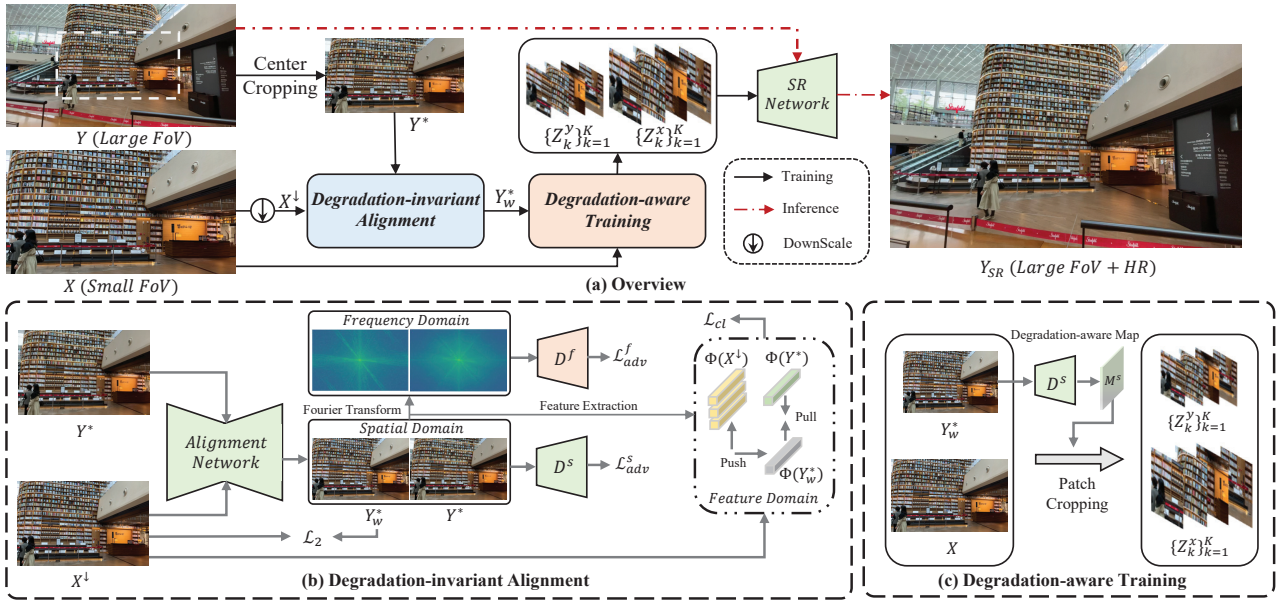


Figure 4. Pipeline of our proposed method. (a) Overview of ZeDuSR, which learns an image-specific SR model on the dual-lens pair at test time. (b) Degradation-invariant alignment, which constrains the alignment process on the spatial, frequency, and feature domains leveraging the adversarial and contrastive learning. (c) Degradation-aware training, which calculates the probability of each location for patch cropping according to a degradation-aware map for reducing the contribution of degradation-variant regions.

With our proposed degradation-invariant alignment method, we then warp the LR image Y^* toward $X^{\downarrow}$ to obtain the aligned LR-HR image pair (Y_w^*, X) with $X^{\downarrow}$ as a bridge[2]. After that, through our designed degradation-aware training strategy, an image-specific SR network is trained with properly selected patches from the aligned image pair. During the inference stage, we apply the SR network to the full wide-angle image Y and generate the HR and large FoV image Y_{SR} beyond the capability of the original device.

4.1. Degradation-invariant Alignment

To warp the LR image Y^* to the downsampled HR image $X^{\downarrow}$, we adopt the photometric consistency loss for training an unsupervised alignment network, formulated as

$$\mathcal{L}_2 = \|X^{\downarrow} - Y_w^*\|_2. \tag{2}$$

However, if we only apply the photometric consistency loss here, the warped LR image Y_w^* would suffer from additional frequency information that is induced by the predefined degradation in $X^{\downarrow}$ (*i.e.*, bilinear here) and lose the realistic degradation in Y^*, as analyzed in Section 3.2. Therefore, to keep the realistic degradation during alignment, we propose to constrain the alignment process in spatial, frequency, and feature domains by leveraging adversarial and contrastive learning, as shown in Figure 4(b).

Spatial adversarial loss. Adversarial learning has been successfully applied to learn data distribution [1,44], which consists of a generator and a discriminator. The generator can learn indistinguishable distribution by fooling the discriminator. Based on this idea, we apply an adversarial loss to preserve the distribution consistency between Y^* and Y_w^* in the spatial domain. Specifically, we take the alignment network as the generator and employ a patch-level discriminator $D^s(\cdot)$ to distinguish the local spatial distribution, since the aligned pair will be cropped to patches for training the SR network. The spatial adversarial loss is denoted as

$$\mathcal{L}_{adv}^s = -\mathbb{E}_{Y_w^*}[\log(D^s(Y_w^*))], \tag{3}$$

and the loss for the patch-level discriminator is in a symmetrical form, denoted as

$$\begin{aligned}\mathcal{L}_{disc}^s = &-\mathbb{E}_{Y^*}[\log(D^s(Y^*))] \\ &-\mathbb{E}_{Y_w^*}[\log(1 - D^s(Y_w^*))].\end{aligned} \tag{4}$$

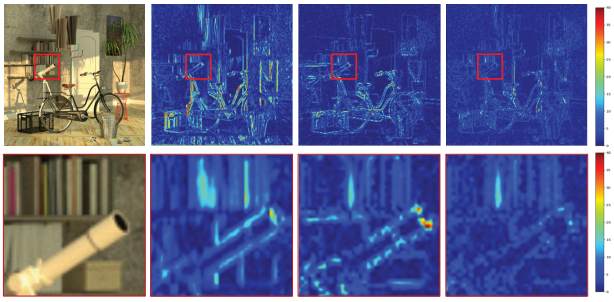
As such, the alignment process is forced to keep the spatial distribution consistency between Y^* and Y_w^*, and thus keep the realistic degradation in Y^*.

Frequency adversarial loss. We also apply an adversarial loss in the frequency domain to avoid introducing additional frequency information in the warped LR image Y_w^*. Specifically, we employ an image-level discriminator $D^f(\cdot)$ to distinguish the global frequency distribution. To this end, we transform Y^* and Y_w^* to the frequency domain by the Fourier Transform. The frequency image can be converted to the amplitude spectrum, formulated as

$$\mathcal{A}[\cdot] = \left[R^2[\cdot] + I^2[\cdot]\right]^{1/2}, \tag{5}$$

where $R^2[\cdot]$ and $I^2[\cdot]$ represent the real and imaginary parts of the frequency image, respectively. Since the amplitude spectrum represents texture information [15,50], we utilize

[2]Warping HR toward LR will result in heavy information loss and limited performance, please see ablation in the supplement.



"ideal" LR Warped LR (SIFT) Warped LR (L_2) Warped LR (Ours)

Figure 5. Error maps between the "ideal" LR image and results from different warping methods. The "ideal" LR image is generated from the HR telephoto image under the same realistic degradation as the wide-angle LR image (only available for simulation).

it to formulate the frequency adversarial loss as

$$\mathcal{L}^f_{adv} = -\mathbb{E}_{Y^*_w}[\log(D^f(\mathcal{A}[Y^*_w]))], \tag{6}$$

and the loss for the image-level discriminator is defined as

$$\begin{aligned}\mathcal{L}^f_{disc} = &-\mathbb{E}_{Y^*}[\log(D^f(\mathcal{A}[Y^*]))] \\ &-\mathbb{E}_{Y^*_w}[\log(1-D^f(\mathcal{A}[Y^*_w]))].\end{aligned} \tag{7}$$

As such, the alignment process is forced to avoid introducing additional frequency information in Y^*_w, and thus keeps the realistic degradation in Y^*.

Feature contrastive loss. Recent works adopt contrastive learning to abstract the representation of images for distinguishing the degradation in the feature domain [20, 42]. Inspired by this idea, we further exploit the feature representation of images to constrain the alignment process. Specifically, we utilize contrastive learning to ensure the warped LR image Y^*_w is pulled closer to the original LR image Y^* and pushed far away from the downsampled HR image $X^{\downarrow}$, in the feature domain realized by the commonly used VGG feature extractor [37]. The feature contrastive loss can be formulated as

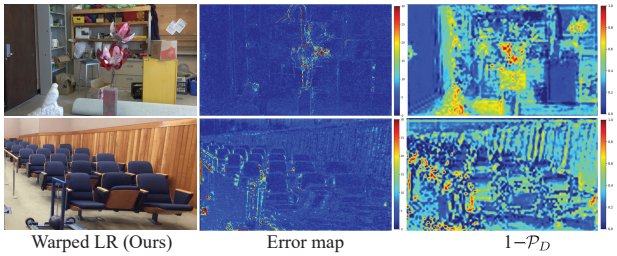
$$\mathcal{L}_{cl} = \frac{\sum_{i=1}^{n}\|\Phi_i(Y^*_w), \Phi_i(Y^*)\|_2}{\sum_{j=1}^{m}\|\Phi_j(Y^*_w), \Phi_j(X^{\downarrow})\|_2}, \tag{8}$$

where $\Phi_i, i = 1, ..., n$ and $\Phi_j, j = 1, ..., m$ represent the i-th and j-th layers from the VGG feature extractor, m and n represent the number of used layers. Different from the perceptual loss [17] that also uses the VGG feature extractor, our contrastive loss adopts the positive-negative samples to constrain the alignment in the feature domain.

Finally, the overall loss function for our unsupervised alignment network is denoted as

$$\mathcal{L}_{align} = \mathcal{L}_2 + \lambda_1\mathcal{L}^s_{adv} + \lambda_2\mathcal{L}^f_{adv} + \lambda_3\mathcal{L}_{cl}, \tag{9}$$

where $\lambda_i, i = 1, 2, 3$ is the weighting factor balancing different loss items. As shown in Figure 5, our proposed alignment method adequately keeps the degradation consistency between the LR images before and after alignment, which plays a key role in training a realistic SR model. The effectiveness of each loss term is validated in Section 7.



Warped LR (Ours) Error map $1-\mathcal{P}_D$

Figure 6. Illustration of the degradation-aware probability map. The error map (between the warped LR image and the "ideal" LR image) indicates regions with degradation variance, which is largely consistent with the inversed probability map.

4.2. Degradation-aware Training

The above alignment process, although designed to be degradation-invariant, cannot be perfect in practice. Therefore, we design a degradation-aware training strategy to reduce the contribution of degradation-variant regions during the training of the image-specific SR network.

Specifically, we generate a probability map for patch cropping during training, by reusing the output map of the patch-level discriminator in the alignment process. Each location of the output map indicates the similarity of the warped patch to the original patch distribution. The larger the discriminator outputs, the lower the possibility that the warped result loses the realistic degradation. By regarding this output map as a degradation-aware map M^s, we then normalize M^s and upsample it (by the bilinear operator) to the same resolution as the warped LR view Y^*_w, and obtain a probability map $\mathcal{P}_D$, denoted as

$$\mathcal{P}_D = (M^s/(sum(M^s)))^{\uparrow}. \tag{10}$$

According to this probability map, the degradation-invariant patches are assigned with a higher probability during training, while the contribution of degradation-variant patches is reduced, as shown in Figure 6. We optimize the SR network with selected LR-HR patches ($\{Z^y_k\}^K_{k=1}$ and $\{Z^x_k\}^K_{k=1}$) using the L_2 loss, formulated as

$$\mathcal{L}_{sr} = \sum_{k=1}^{K}\|Z^x_k - Z^y_k\|_2, \tag{11}$$

where K represents the number of patches. In this way, we can fully exploit the limited training data and reduce the influence of degradation-variant regions. The effectiveness of our strategy is validated in Section 7.

5. Experiments on Synthesized Data

Datasets. To simulate the dual-lens configuration on mobile devices, we adopt a light field dataset (HCI new [12], two views selected, with a small baseline) and a stereo dataset (Middlebury2021 [35], with a relatively large baseline) for quantitative evaluation, where the HR ground-truth of the synthesized wide-angle view is available. To generate the dual-lens image pair, we apply three groups of image-specific degradation kernels, *i.e.*, isotropic

Table 1. Quantitative comparisons with SISR, RefSR, and dual-lens SR for 2× and 4× SR on synthesized dual-lens data with image-specific degradation (IG and AG). PSNR/SSIM (the higher, the better) are adopted for the evaluation of reconstruction accuracy. **Red**, blue, and orange indicate the best, second best, and third best performance, respectively.

Method	HCI_new				Middlebury2021			
	IG		AG		IG		AG	
	2× scale	4× scale	2× scale	4× scale	2× scale	4× scale	2× scale	4× scale
Bicubic	29.45/0.8119	27.62/0.7168	28.27/0.7505	27.33/0.7003	33.43/0.9409	30.87/0.9001	32.44/0.9267	30.33/0.8911
RCAN [52]	30.15/0.8371	28.23/0.7415	28.99/0.7670	28.00/0.7191	34.11/0.9465	31.84/0.9084	33.17/0.9326	31.30/0.9053
CSNLN [31]	30.24/0.8376	28.25/0.7379	28.94/0.7682	27.97/0.7172	34.23/0.9487	31.89/0.9136	33.10/0.9366	31.39/0.9040
SwinIR [22]	30.19/0.8329	28.27/0.7405	28.84/0.7652	27.90/0.7140	34.28/0.9492	31.96/0.9144	33.05/0.9361	31.27/0.9022
ZSSR [36]	30.29/0.8236	27.75/0.7186	28.63/0.7580	27.42/0.7041	33.93/0.9453	30.92/0.9021	32.91/0.9331	30.49/0.8925
KernelGAN [1]	-	-	29.78/0.8042	28.36/0.7358	-	-	33.65/0.9399	31.83/0.9032
TTSR [46]	-	28.03/0.7367	-	27.67/0.7082	-	31.39/0.9045	-	30.97/0.8991
MASA [25]	-	28.32/0.7498	-	28.05/0.7212	-	32.01/0.9094	-	31.42/0.9011
DCSR [43]	30.40/0.8306	28.38/0.7440	29.22/0.7627	27.99/0.7176	34.29/0.9434	32.02/0.9073	33.18/0.9382	31.36/0.8992
SelfDZSR [53]	29.86/0.8382	27.91/0.7297	28.97/0.7943	27.60/0.7244	33.91/0.9442	31.32/0.9132	32.96/0.9405	31.03/0.9033
DCSR+SRA [43]	30.61/0.8424	28.56/0.7486	29.44/0.7910	28.18/0.7317	34.37/0.9468	32.17/0.9124	33.38/0.9404	31.59/0.9037
ZeDuSR	31.01/0.8529	28.87/0.7536	30.02/0.8146	28.66/0.7420	34.78/0.9553	32.41/0.9117	33.79/0.9421	31.76/0.9019
ZeDuSR*	**31.17/0.8594**	**29.25/0.7601**	**30.23/0.8183**	**29.09/0.7483**	**34.89/0.9571**	**32.77/0.9212**	**33.94/0.9450**	**32.42/0.9141**

Table 2. Quantitative comparisons with blind SR for 2× and 4× SR on synthesized dual-lens data with image-specific degradation (IG and IG_JPEG). PSNR/SSIM are adopted for the evaluation of reconstruction accuracy.

Method	HCI_new				Middlebury2021			
	IG		IG_JPEG		IG		IG_JPEG	
	2× scale	4× scale	2× scale	4× scale	2× scale	4× scale	2× scale	4× scale
Bicubic	29.45/0.8119	27.62/0.7168	29.42/0.7839	27.21/0.6881	33.43/0.9409	30.87/0.9001	33.02/0.9277	30.41/0.8822
DANv1 [27]	30.29/0.8569	28.43/0.7534	29.07/0.7747	27.19/0.6917	34.44/0.9519	32.22/0.9159	33.10/0.9182	30.45/0.8720
DANv2 [28]	30.16/0.8539	28.48/0.7586	28.82/0.7721	27.28/0.6911	34.36/0.9516	32.29/0.9150	32.97/0.9169	30.56/0.8695
DCLS [26]	30.63/0.8651	28.68/0.7524	29.21/0.7841	27.45/0.6928	34.72/0.9528	32.46/0.9173	33.38/0.9250	30.71/0.8725
ZeDuSR	31.01/0.8529	28.87/0.7536	29.98/0.8006	27.82/0.7035	34.78/0.9553	32.41/0.9171	33.63/0.9316	30.97/0.8805
ZeDuSR*	**31.17/0.8594**	**29.25/0.7601**	**30.06/0.8032**	**28.17/0.7179**	**34.89/0.9571**	**32.77/0.9212**	**33.79/0.9332**	**31.57/0.8929**

and anisotropic Gaussian downsampling (IG and AG) and isotropic Gaussian downsampling with slight JPEG compression (IG_JPEG)[3], on one of the two views (acting as wide-angle), while the other view acts as telephoto after center cropping. Details of synthesized data generation and degradation simulation are in the supplement.

Implementation Details. For the alignment network, we modify FlowNet-S [14] by decreasing the convolution layers to suit the limited training data. We use the RCAN [52] backbone as the image-specific SR network. The embodiments of the alignment and SR networks can be replaced, the ablation studies are in Section 7. More implementation details are in the supplement.

Comparison Methods. We compare ZeDuSR with several representative methods including five categories: 1) non-blind single-image SR (SISR): RCAN [52], CSNLN [31], and SwinIR [22], 2) blind SISR with a kernel pool: DANv1 [27], DANv2 [28], and DCLS [26], 3) zero-shot SISR with kernel prediction: ZSSR [36] and KernelGAN [1], 4) reference-based SR (RefSR): TTSR [46] and MASA [25], 5) dual-lens SR: DCSR [43], DCSR+SRA [43], and SelfDZSR [53]. SISR methods use only the wide-angle image during training and inference, while RefSR and dual-lens SR methods use the wide-angle/telephoto image pair. These methods are reproduced with the best possible training configurations following their original papers. Besides, we provide an updated version of ZeDuSR to demonstrate its promising potential (denoted as ZeDuSR*), where the parameters of the image-specific SR network are initialized by a pretrained model with bicubic degradation. Note that, still, no extra dual-lens data is required during training.

Quantitative Comparison. We compare ZeDuSR with non-blind SISR, zero-shot SISR, RefSR, and dual-lens SR under image-specific degradation IG and AG in Table 1. The methods using external training data with bicubic degradation (*e.g.*, SwinIR [22], MASA [25], and DCSR [43]) inevitably face the domain gap caused by the degradation variance between training and inference. Zero-shot SISR methods (*e.g.*, KernelGAN [1]) may meet difficulty for degradation kernel estimation. Self-supervised dual-lens SR methods (DCSR+SRA [43] and SelfDZSR [43]) still assume consistent degradation on the external training data from the target device, which limits their generalization capabilities. It can be seen that ZeDuSR shows superior performance over the previous methods in most cases, thanks to the image-specific degradation assumption. Meanwhile, ZeDuSR* achieves further improved results, surpassing the previous methods by a large margin. We compare ZeDuSR with blind SISR methods under image-specific degradation IG and IG_JPEG in Table 2,

[3]JPEG is commonly used on mobile devices, which is signal-dependent and thus image-specific degradation, even at fixed compression ratio.

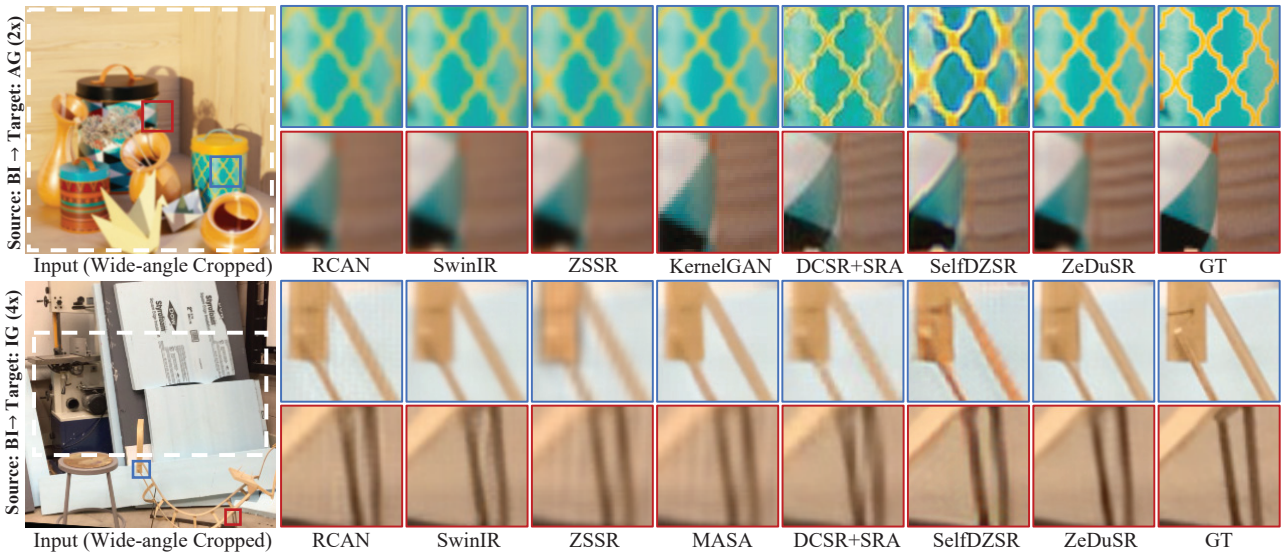


Figure 7. Visual comparisons on synthesized data. The white dotted box indicates the overlapped FoV. Top: "Origami" from HCI_new. Bottom: "Bandsaw1" from Middlebury2021.

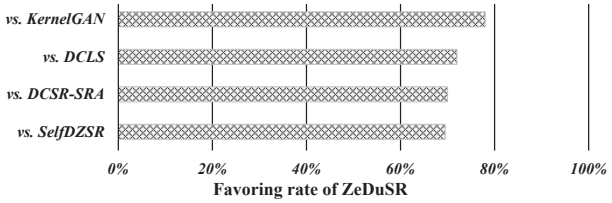


Figure 8. User study on the real-world dual-lens SR results.

Table 3. Ablation on the loss function and training strategy.

L_{adv}^{s}	L_{adv}^{f}	L_{cl}	DaTS	PSNR	SSIM
✗	✗	✗	✗	30.48	0.8362
✓	✗	✗	✗	30.59	0.8429
✓	✓	✗	✗	30.72	0.8452
✓	✓	✓	✗	30.89	0.8512
✓	✓	✓	✓	31.01	0.8529

where the blind SISR methods are trained using external data with a kernel pool of AG degradation. As can be seen, ZeDuSR achieves better results overall since the realistic degradation is not within the kernel pool of blind SISR.

Qualitative Comparison. Examples of qualitative comparison between ZeDuSR and other methods are shown in Figure 7. It can be observed that our method recovers more accurate and reliable details, while others suffer from blurry or unrealistic artifacts due to the degradation variance.

6. Real-world Experiments

Dataset. We conduct experiments on real-world datasets captured by off-the-shelf smartphones, including CameraFusion [43] (by iPhone11) and RealMCVSR [19] (by iPhone12). The former supports $2\times$ dual-lens SR while the latter supports both $2\times$ and $4\times$ dual-lens SR. More details of the datasets are in the supplement.

Statistic Evaluation. Since the HR ground-truth image is NOT available on the real-world dual-lens data, we perform a user study on the real-world results generated by representative methods. We provide the users with anonymous pair comparisons (ZeDuSR vs. the other method) and ask them to select the one with higher quality. We collect 1200 votes from 30 users and the statistics are summarized in Figure 8. The user study verifies the superiority of our method on real-world data. We also provide non-reference evaluations in the supplement.

Visual Comparison. Examples of real-world visual results are shown in Figure 9. It can be observed that ZeDuSR reconstructs higher-fidelity outputs within and outside the overlapped FoV, by leveraging the implicit modeling of the image-specific degradation, where more realistic textures and sharper edges are recovered. More real-world results are provided in the supplement.

7. Ablation Study

Loss Function and Training Strategy. To investigate the effectiveness of our proposed loss function for degradation-invariant alignment and the degradation-aware training strategy (DaTS), we conduct an ablation study on the HCI_new dataset with IG degradation for $2\times$ SR. As can be seen in Table 3, the PSNR gradually increases by adding the three loss terms to the L_2 loss when training the alignment network. If the image-specific SR network is trained with DaTS, the PSNR further increases.

Alignment and SR backbones. To investigate the impact of the embodiments for the alignment network and the image-specific SR network, we take SPyNet [34] and PWCNet [39] as the alternative alignment networks, while CSNLN [31] and SwinIR [22] are selected as the alterna-

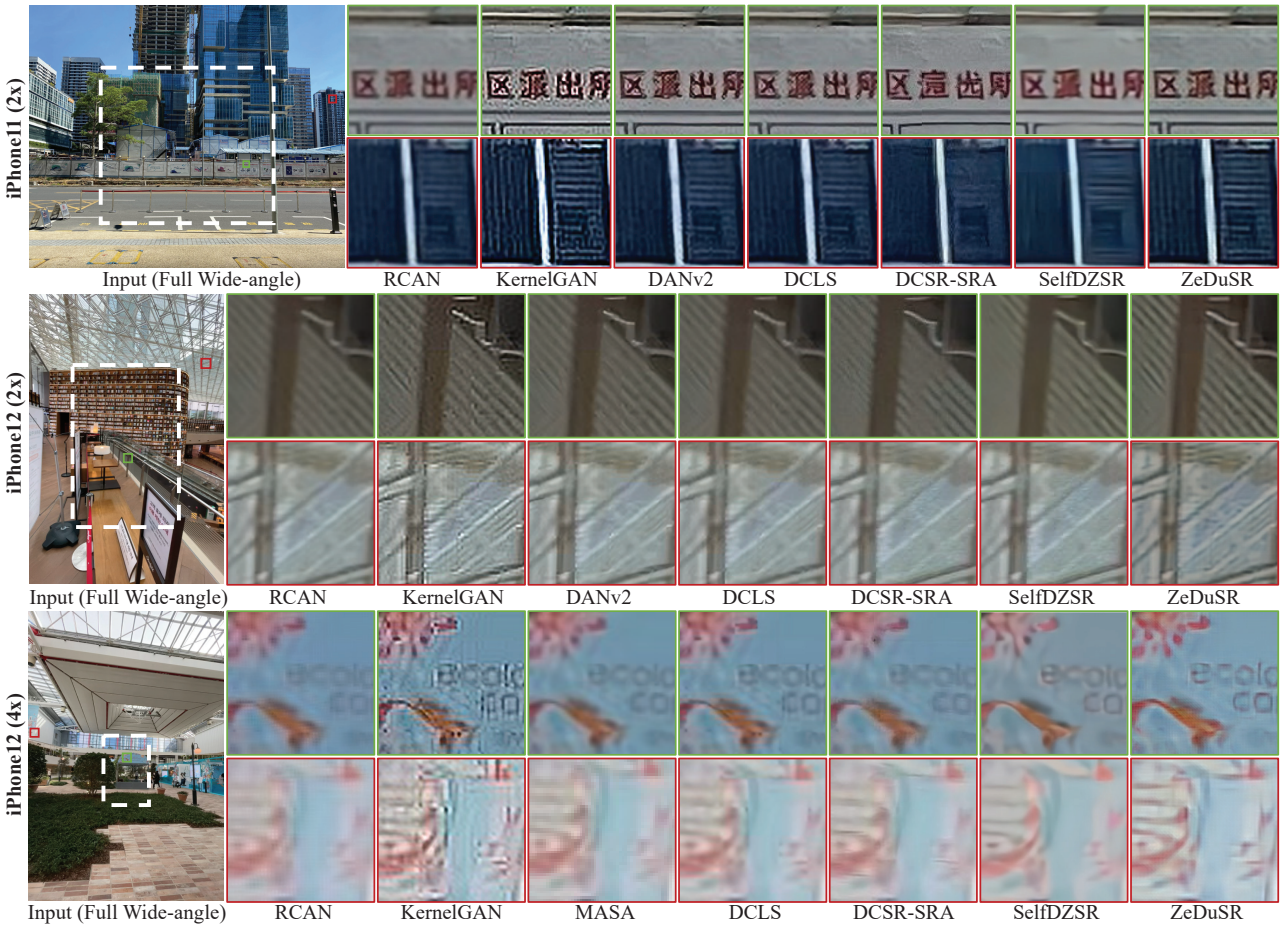


Figure 9. Visual comparisons on real-world data. The white dotted box indicates the overlapped FoV. **More results are in the supplement.**

Table 4. Ablation on the alignment backbone.

Alignment Network	IG	AG
FlowNet-S [8]	31.01/0.8529	30.02/0.8146
SPyNet [34]	31.05/0.8533	29.97/0.8139
PWCNet [39]	29.98/0.8515	30.06/0.8159

Table 5. Ablation on the SR backbone.

SR Network	IG	AG
RCAN [52]	30.15/0.8371	28.99/0.7670
RCAN [52] + ZeDuSR	31.01/0.8529	30.02/0.8146
CSNLN [31]	30.24/0.8376	28.94/0.7682
CSNLN [31] + ZeDuSR	31.03/0.8516	29.97/0.8135
SwinIR [22]	30.19/0.8329	28.84/0.7652
SwinIR [22] + ZeDuSR	31.04/0.8531	29.91/0.8128

tive SR networks. We retrain our method on the HCI_new dataset with IG and AG degradation for 2× SR. As can be seen in Table 4 and Table 5, the performances of different embodiments are close, which demonstrates the robustness of ZeDuSR for different backbones.

8. Conclusion

In this paper, we present a zero-shot learning solution for dual-lens SR (ZeDuSR), which learns an image-specific SR model with the single dual-lens pair at test time. Specifically, we propose a degradation-invariant alignment method to generate an aligned LR-HR image pair for training the SR model while keeping the realistic degradation, along with a degradation-aware training strategy to effectively exploit the information within the highly limited training data. Experiments on synthesized and real-world datasets demonstrate the superiority of ZeDuSR over existing solutions.

Similar to zero-shot SISR methods, ZeDuSR has the drawback of online training, *i.e.*, a long inference time (see supplement). This problem can be alleviated by initializing the parameters with a pretrained model as in ZeDuSR*, or it will no longer be a problem if cloud computing is enabled on mobile devices. We believe ZeDuSR moves closer to addressing realistic SR on the widely available dual-lens devices, which would benefit downstream applications that require HR and large FoV inputs.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *NeurIPS*, 2019. 1, 2, 4, 6

[2] 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 *ICCV*, 2019. 2

[3] Chang Chen, Xinmei Tian, Zhiwei Xiong, and Feng Wu. Udnet: Up-down network for compact and efficient feature representation in image super-resolution. In *ICCVW*, 2017. 2

[4] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, 2019. 2, 3

[5] Xihao Chen, Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Yueyi Zhang, and Zheng-Jun Zha. Degradation-agnostic correspondence from resolution-asymmetric stereo. In *CVPR*, 2022. 3

[6] Xi Cheng, Zhenyong Fu, and Jian Yang. Zero-shot image super-resolution with depth guided internal degradation learning. In *ECCV*, 2020. 1

[7] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *CVPR*, 2021. 2

[8] 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 *ICCV*, 2015. 8

[9] Yuanbiao Gou, Boyun Li, Zitao Liu, Songfan Yang, and Xi Peng. Clearer: Multi-scale neural architecture search for image restoration. *NeurIPS*, 2020. 2

[10] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 2

[11] Baisong Guo, Xiaoyun Zhang, Haoning Wu, Yu Wang, Ya Zhang, and Yan-Feng Wang. Lar-sr: A local autoregressive model for image super-resolution. In *CVPR*, 2022. 2

[12] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, 2016. 2, 5

[13] 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 *CVPR*, 2022. 3

[14] 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 *CVPR*, 2017. 3, 6

[15] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, 2021. 4

[16] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, 2021. 3

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[18] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 2022. 2

[19] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 2022. 7

[20] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, 2022. 5

[21] Jiacheng Li, Chang Chen, Zhen Cheng, and Zhiwei Xiong. Mulut: Cooperating multiple look-up tables for efficient image super-resolution. In *ECCV*, 2022. 2

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR*, 2021. 6, 7, 8

[23] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *ICCV*, 2021. 2

[24] Yucheng Liu and Buyue Zhang. Photometric alignment for surround view camera system. In *ICIP*, 2014. 3

[25] 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 *CVPR*, 2021. 3, 6

[26] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *CVPR*, 2022. 6

[27] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. In *NeurIPS*, 2020. 6

[28] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. End-to-end alternating optimization for blind super resolution. *arXiv preprint arXiv:2105.06878*, 2021. 6

[29] Sai Kumar Reddy Manne, BH Prasad, and KS Rosh. Asymmetric wide tele camera fusion for high fidelity digital zoom. In *ICCVIP*, 2019. 3

[30] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, 2021. 2

[31] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 6, 7, 8

[32] Zhihong Pan, Baopu Li, Dongliang He, Mingde Yao, Wenhao Wu, Tianwei Lin, Xin Li, and Errui Ding. Towards bidirectional arbitrary image rescaling: Joint optimization and cycle idempotence. In *CVPR*, 2022. 2

[33] Seoyoung Park, Byeongho Moon, Seonhee Park, Seungyong Ko, Soohwan Yu, and Joonki Paik. Brightness and color correction for dual camera image registration. In *ICCE-Asia*, 2016. 3

[34] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 7, 8

[35] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 2014. 2, 5

[36] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, 2018. 1, 2, 6

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[38] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *CVPR*, 2020. 2

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 7, 8

[40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 1

[42] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, 2021. 5

[43] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *ICCV*, 2021. 1, 2, 3, 6, 7

[44] 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 *ECCVW*, 2018. 4

[45] Ruikang Xu, Zeyu Xiao, Mingde Yao, Yueyi Zhang, and Zhiwei Xiong. Stereo video super-resolution via exploiting view-temporal correlations. In *ACM MM*, 2021. 3

[46] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 3, 6

[47] Mingde Yao, Zhiwei Xiong, Lizhi Wang, Dong Liu, and Xuejin Chen. Spectral-depth imaging with deep learning based reconstruction. *Optics Express*, 27(26):38312–38325, 2019. 3

[48] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, 2018. 2

[49] Zongsheng Yue, Qian Zhao, Jianwen Xie, Lei Zhang, Deyu Meng, and Kwan-Yee K. Wong. Blind image super-resolution with elaborate degradation modeling on noise and kernel. In *CVPR*, 2022. 2

[50] Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *CVPR*, 2022. 4

[51] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, 2019. 2

[52] 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. 6, 8

[53] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, Yunjin Chen, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual zoomed observations. In *ECCV*, 2022. 1, 2, 3, 6

[54] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, 2019. 3

[55] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans. Circuits Syst. Video Technol.*, 32(3):1186–1196, 2022. 3

[56] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, 2022. 3

[57] 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 *ECCV*, 2018. 3