Benchmark Dataset and Effective Inter-Frame Alignment for Real-World Video Super-Resolution

Ruohao Wang[1], Xiaohui Liu[1], Zhilu Zhang[1], Xiaohe Wu[1](✉), Chun-Mei Feng[2], Lei Zhang[3], Wangmeng Zuo[1]

[1]Harbin Institute of Technology, China
[2]Institute of High Performance Computing, A*STAR, Singapore
[3]The Hong Kong Polytechnic University, China

{rhwangHIT, cszlzhang}@outlook.com, {lxh720199, xhwu.cpsl.hit, strawberry.feng0304}@gmail.com
cslzhang@comp.polyu.edu.hk, wmzuo@hit.edu.cn

Abstract

Video super-resolution (VSR) aiming to reconstruct a high-resolution (HR) video from its low-resolution (LR) counterpart has made tremendous progress in recent years. However, it remains challenging to deploy existing VSR methods to real-world data with complex degradations. On the one hand, there are few well-aligned real-world VSR datasets, especially with large super-resolution scale factors, which limits the development of real-world VSR tasks. On the other hand, alignment algorithms in existing VSR methods perform poorly for real-world videos, leading to unsatisfactory results. As an attempt to address the aforementioned issues, we build a real-world ×4 VSR dataset, namely MVSR4×, where low- and high-resolution videos are captured with different focal length lenses of a smartphone, respectively. Moreover, we propose an effective alignment method for real-world VSR, namely EAVSR. EAVSR takes the proposed multi-layer adaptive spatial transform network (MultiAdaSTN) to refine the offsets provided by the pre-trained optical flow estimation network. Experimental results on RealVSR and MVSR4× datasets show the effectiveness and practicality of our method, and we achieve state-of-the-art performance in real-world VSR task. The dataset and code will be available at https://github.com/HITRainer/EAVSR.

1. Introduction

In comparison with single image super-resolution (SR) that reconstructs a high-resolution (HR) image from its low-resolution (LR) counterpart, video super-resolution (VSR) can further exploit valuable information from neighboring frames to recover richer textures and more details. Thus, the inter-frame temporal alignment plays a significant role in VSR. Driven by the development of convolutional neural
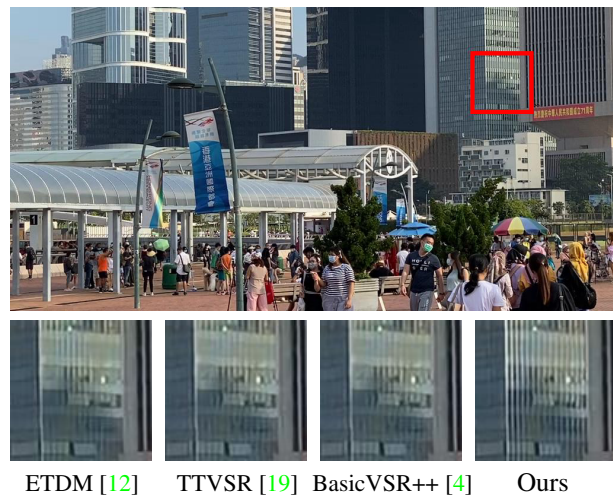

ETDM [12]　　TTVSR [19]　　BasicVSR++ [4]　　Ours

Figure 1. Super-resolution results on a real-world video. In comparison with other VSR algorithms (*i.e.*, ETDM [12], TTVSR [19] and BasicVSR++ [4]), our method restores more textures. Please zoom in for more details.

networks (CNNs), especially the alignment methods [7, 22, 40], tremendous progress has been made in VSR task.

However, it is challenging for existing VSR methods to be effective in real-world applications. On the one hand, most VSR methods are trained only with synthetic (*e.g.*, bicubic downsampling) LR videos. Although RealBasicVSR [5] adopts second-order order degradation model [30] and video compression to synthesize LR frames, its generalization capability on real-world LR videos is still limited as real degradations are more complex and changeable. A natural solution is to capture real-world video pairs and take them to train the network. Yang *et al*. [34] has proposed a real-world VSR dataset with ×2 scale factor, but larger scale factor (*e.g.*, ×4) datasets are wanting. In addition, it is also more difficult to collect and process VSR

datasets with larger scale-factor.

On the other hand, alignment algorithms in existing VSR methods perform poorly for real-world videos, leading to unsatisfactory results. For example, TOF [33] and BasicVSR [3] estimate the offset between neighboring frames by optical flow network [22] pre-trained on clean images. When facing low-quality real-world frames, the methods may not predict accurate offsets. EDVR [29] and TDAN [27] introduce deformable convolution network (DCN) [7] for inter-frame alignment. Although such alignment methods are task-oriented and have more flexibility, it is easy to lead to the instability of network training [2]. Chan *et al.* [4] combine the optical flow network and DCN together, but it is still not effective enough for aligning real-world frames and can be further improved.

As an attempt to address the aforementioned issues, we build a new real-world video super-resolution dataset with ×4 scale factor by mobile phone (*i.e.*, Huawei P50 Pro), named MVSR4×. The dual-view function in the smartphone is leveraged to take video frames with different focal length lenses with the same frequency almost synchronously. We collect the videos in different moving speeds, lightness, and scenes to guarantee data diversity. In addition, it is inevitable that position misalignment and color inconsistency exist between images with different focal lengths. To mitigate the adverse effect of these discrepancies on VSR model learning, we apply color correction and position alignment algorithms [24, 25] not only in data pre-processing but also in the training phase.

Furthermore, we propose an effective alignment method for real-world VSR, namely EAVSR, where a multi-layer adaptive spatial transform network (MultiAdaSTN) is proposed to refine inter-frame offsets. Specifically, MultiAdaSTN consists of ResflowNet and DeformNet. Considering that the pre-trained optical flow network can provide a rough offset for neighboring frames, we utilize it to estimate a basic flow preliminarily. Noted that the basic flow is not completely accurate, and only taking DCN [7] to refine it (like BasicVSR++ [4]) may be inadequate as the optimization of DCN is hard. Instead, we first compute a residual flow relative to the basic flow by ResflowNet, which introduces AdaSTN [39] and works in a pyramid coarse-to-fine manner. The combination of basic and residual flow can be considered as a more precise optical flow. Then DeformNet is present to refine the optical flow to get ultimate inter-frame offsets. Different from DCN, DeformNet utilizes the modified AdaSTN (*i.e.*, AdaSTN v2) to predict the offsets and masks of deformable convolution v2 [40].

The temporal propagation of our EAVSR adopts the bidirectional scheme of BasicVSR [3]. And we also introduce second-order grid propagation proposed by BasicVSR++ [4] into EAVSR to get EAVSR+. Experiments are conducted on RealVSR [34] and MVSR4× datasets.

The results show the effectiveness and practicality of our proposed method, and it achieves state-of-the-art performance in real-world VSR task.

To sum up, the main contributions of this work include:

- We construct a benchmark dataset MVSR4×, which compensates for the lack of paired real-world video datasets with large scale factors.

- We propose an effective alignment method EAVSR towards real-world data, where a multi-layer adaptive spatial transformer network (MultiAdaSTN) is designed elaborately to refine the inter-frame offsets.

- Experiments on RealVSR dataset and MVSR4× datasets show that our method achieves state-of-the-art performance in real-world VSR task.

2. Related Work

2.1. Degradation Model

Most previous VSR methods [2–4, 8, 10, 11, 13, 18, 23, 26–28, 33] focus on exploring inter-frame alignment and restoration networks based on synthetic datasets (*e.g.*, REDS [20] and Vimeo-90K [33]), where LR videos are generated by bicubic down-sampling HR ones. While significant performance has been gained on the synthetic datasets, the generalized results on real-world data are very limited due to the degradation gap. For bridging this gap, RealBasicVSR [5] makes use of a second-order order degradation model [30] and video compression to synthesize LR frames, but its generalization capability on real-world LR videos is still unsatisfactory as real degradations are more complex and changeable. Actually, to build VSR models for real scenarios, collecting real-world paired LR-HR videos is essential. Recently, Yang *et al.* [34] build a ×2 real-world video super-resolution dataset RealVSR by capturing paired LR-HR video sequences using the multi-camera system of iPhone 11 Pro Max. In this work, to promote the development of real-world VSR, we construct a ×4 real-world video super-resolution dataset, namely MVSR4×.

2.2. Inter-Frame Alignment

Frame alignment is of great importance to use the complementary information across frames effectively in video restoration [3, 21, 32, 33]. Optical flow networks and their improved variants are extensively studied and deployed for motion estimation in VSR. TOFlow [33] proposes a task-oriented flow module, which can be jointly trained with a video processing network. OFRNet [28] introduces an optical flow reconstruction network to infer HR optical flow in a coarse-to-fine manner for performing motion compensation. However, optical flow networks can't deal with occlusions and sometimes make inaccurate predictions, impairing the restoration performance.
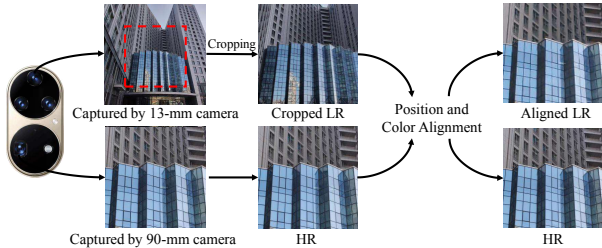
Figure 2. Pre-processing pipeline of our MVSR4× dataset.

Table 1. Comparison of existing VSR datasets.

Dataset	Real-World	Ground-Truth	Aligned	Scale Factor
Vimeo-90K [33]	✗	✓	✓	4
REDS [20]	✗	✓	✓	4
VideoLQ [5]	✓	✗	-	-
RealMCVSR [17]	✓	✗	✗	4
RealVSR [34]	✓	✓	✓	2
MVSR4×	✓	✓	✓	4

To alleviate the above issues, recent researchers aim to exploit deformable convolution network (DCN) [7] to align frames of videos. For instance, TDAN [27] leverages DCN to perform alignment at the feature level instead of explicit motion estimation or image warping. Inspired by TDAN, EDVR [29] deploys a pyramid cascading and deformable convolutions module to tackle the large motion. Although DCN-based alignment is task-oriented with more flexibility, it shows poor training stability. Recently, BasicVSR++ [4] proposes a flow-guided deformable alignment, but it is not effective enough for aligning neighboring frames from the real world.

Except for explicitly deploying the inter-frame alignment module, many algorithms use an indirect approach to fuse inter-frame information. FSTRN [18] attempts to use 3D convolution networks to transfer temporal information. ETDM [12] computes the temporal difference between frames, divides those pixels into two subsets according to the level of difference, and uses different encoder networks to learn the distribution of these two differences separately. In addition, with the development of Transformer, some methods achieve feature fusion by calculating the similarity of patches between frames. TTVSR [19] leverages the self-attention module to calculate and focus on the relevant spatio-temporal trajectories. MANA [35] designs a robust non-local attention mechanism that allows frames existing misalignment. However, these methods are generally computationally expensive.

3. Real-World ×4 VSR Dataset

Existing VSR methods are mostly trained on synthetic data where LR sequences are obtained by some sim-

ple degradations (*e.g.*, bicubic downsampling) of HR sequences. However, these degradations can not fully reflect the degradation of real-world videos, and the VSR methods hardly generate high-quality results in real scenarios. A feasible solution is to construct real-world datasets to train VSR models. Some relevant datasets have been proposed, the comparison can be seen in Table 1. For instance, Chan *et al.* [5] collect the real-world VSR dataset VideoLQ, which only has LR sequences but not their HR counterparts. While RealMCVSR [17] captures paired videos by smartphone lenses with different focal lengths, the data pairs are not aligned. In addition, Yang *et al.* [34] proposed a ×2 real-world aligned VSR dataset. However, it is not sufficient to verify the effectiveness of VSR methods, as the scale factor is low and the degradation is a little weak. Therefore, there is still a high demand for paired, aligned, and larger scale factor VSR datasets from the real world.

In this work, we collect a new real-world VSR dataset with the scale factor of 4 by mobile phone (*i.e.*, Huawei P50 Pro), named MVSR4×. As illustrated in Fig. 2, we use an ultra-wide camera with a 13mm-equivalent (r_u) lens and a telephoto camera with a 90mm-equivalent (r_t) lens to capture LR and HR sequences, respectively. Thanks to the dual-view video function provided by the camera engine, we can easily capture LR and HR videos at the same frequency and almost synchronously. The focal length scale of HR and LR videos is $r_t/r_u \approx 6.9$. For simplicity, we regard it as the dataset with a common scale factor of 4. Thus, we need to crop the LR size to $r_u/(r_t/4) \approx 0.58$ times the original size. This operation can also remove areas with wide-angle distortion. Finally, we utilize the alignment algorithm of RealSR [1] to correct the spatial position and color of LR in an iterative manner.

We select 200 high-quality LR-HR sequences from the captured 500 videos to form the MVSR4× dataset. Each sequence has 100 frames whose frame rate is 30fps and resolution is 1080P. Fig. 3 illustrates some image examples. Moreover, the dataset consists of diverse contents, including slow and fast movement, bright and dark, indoor and outdoor scenes, as shown in Fig. 4. There are 180 sequences used for training and 5 sequences for validation. The remaining 15 sequences are used for testing.

4. Method

In this section, we first give an overview of our pipeline in Sec. 4.1. Then we introduce the proposed alignment method for neighboring frames from the real world in Sec. 4.2 and training loss in Sec. 4.3.

4.1. Overview

The whole VSR pipeline consists of four parts, *i.e.*, encoding model, alignment model, reconstruction model, and upsampling model.
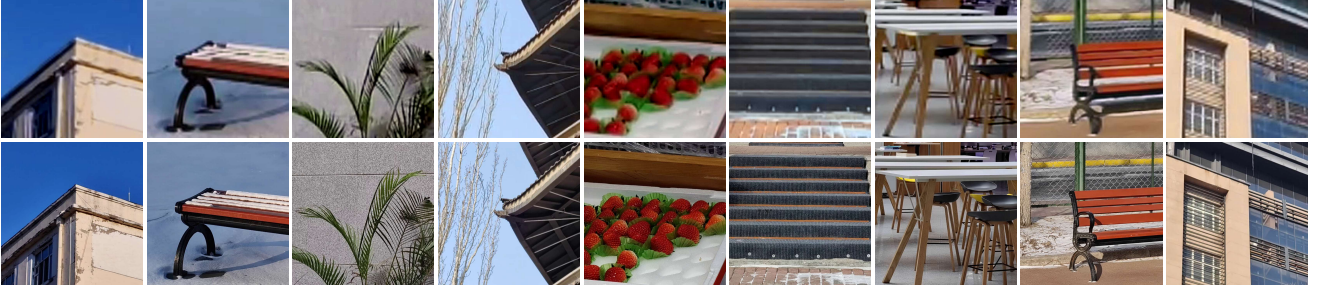
Figure 3. Some images from the MVSR4× dataset. The first row shows examples of LR frames and the second row shows examples of corresponding aligned HR frames.
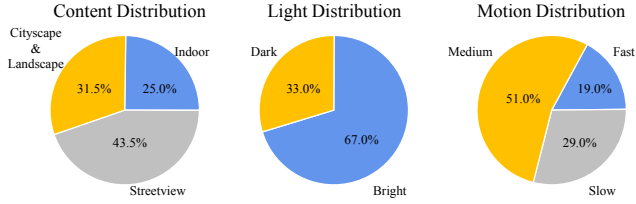

Figure 4. Content and scenarios analysis of our MVSR4× dataset.

Denote by $x \in \mathbb{R}^{T \times H \times W \times C}$ low-resolution videos. When reconstructing the i-th frame x_i, x_i is firstly fed into the encoder model for feature extraction,

$$f_i = \mathcal{E}(x_i), \qquad (1)$$

where $\mathcal{E}$ denotes the encoder and f_i is the output features. Taking forward propagation as an example, the alignment model calculates the offsets between the current frame x_i and the previous frame x_{i-1}, and then warps the reconstruction features h_{i-1} of x_{i-1} to get the aligned features $\bar{h}_i$. The restoration features are recursively generated from the restoration model $\mathcal{R}$, it can be expressed as,

$$h_i = \mathcal{R}(x_i, \bar{h}_i). \qquad (2)$$

Finally, the upsampling module $\mathcal{U}$ is deployed to generate the super-resolution result $\hat{y}_i$,

$$\hat{y}_i = \mathcal{U}(h_i), \qquad (3)$$

where $\hat{y}_i \in \mathbb{R}^{T \times rH \times rW \times C}$ and r denotes the scale factor of super-resolution.

To estimate the effectiveness of the proposed method comprehensively, we select bidirectional [3] and second-order grid [4] as our temporal propagation scheme, named EAVSR and EAVSR+, respectively. Due to the space limitation, the specific structure is introduced in the supplementary material.

4.2. Inter-Frame Ailgnment

Since the pre-trained optical flow network sometimes provides inaccurate results on real-world data and de-

formable convolution [7] is unstable during training. We consider designing a more effective alignment method for real-world VSR. Specifically, we propose a multi-layer adaptive spatial transform network (MultiAdaSTN) to refine inter-frame offsets provided by a pre-trained optical flow network. As shown in Fig. 5, the proposed Multi-AdaSTN contains ResflowNet and DeformNet.

ResflowNet. Given two neighboring frames x_{i-1} and x_i, we first calculate a basic optical flow $\Psi_{i-1 \to i}$ from x_{i-1} to x_i by a pre-trained optical flow network $\mathcal{F}_s$ (e.g., SpyNet [22]),

$$\Psi_{i-1 \to i} = \mathcal{F}_s(x_{i-1}, x_i). \qquad (4)$$

ResflowNet estimates a residual optical flow $\Delta\Psi_{i-1 \to i}$ on the basis of $\Psi_{i-1 \to i}$ in a coarse-to-fine manner. To generate the multi-level features $f_{i;l}$, $f_{i-1;l}$ and flow $\Psi_{i-1 \to i;l}$ at the l-th level, the features and flow at the (l-1)-th level are downsampled by a factor of 2. In this way, we can obtain all l-level features and flow.

Before performing the l-th level, the residual flow $\Delta\Psi_{i-1 \to i;l+1}$ from (l+1)-th level can be acquired. Then the residual and basic flow are added together to generate the l-th coarse flow,

$$\Psi_{i-1 \to i;l} = \Psi_{i-1 \to i;l} + (\Delta\Psi_{i-1 \to i;l+1})_{\uparrow 2}, \qquad (5)$$

where $(\cdot)_{\uparrow s}$ refers to the bilinear upsampling with the scale factor of s. Next, $f_{i-1;l}$ can be warped with $\Psi_{i-1 \to i;l}$ to obtain the preliminary aligned feature $\bar{f}_{i-1;l}$,

$$\bar{f}_{i-1;l} = \mathcal{W}(f_{i-1;l}, \Psi_{i-1 \to i;l}), \qquad (6)$$

where $\mathcal{W}$ denotes warping operation. Following, we utilize the AdaSTN [39] $\mathcal{S}$ (shown in Fig. 6 (a)) to calculate fine residual flow between $\bar{f}_{i-1;l}$ and $f_{i;l}$,

$$\Delta\tilde{\Psi}_{i-1 \to i;l} = \mathcal{S}(f_{i;l}, \bar{f}_{i-1;l}). \qquad (7)$$

And the residual flow $\Delta\Psi_{i-1 \to i;l}$ between $f_{i;l}$ and $f_{i-1;l}$ can be regarded as the sum of residue propagated from (l+1)-th level and calculated in l-th level,

$$\Delta\Psi_{i-1 \to i;l} = (\Delta\Psi_{i-1 \to i;l+1})_{\uparrow 2} + \Delta\tilde{\Psi}_{i-1 \to i;l}. \qquad (8)$$
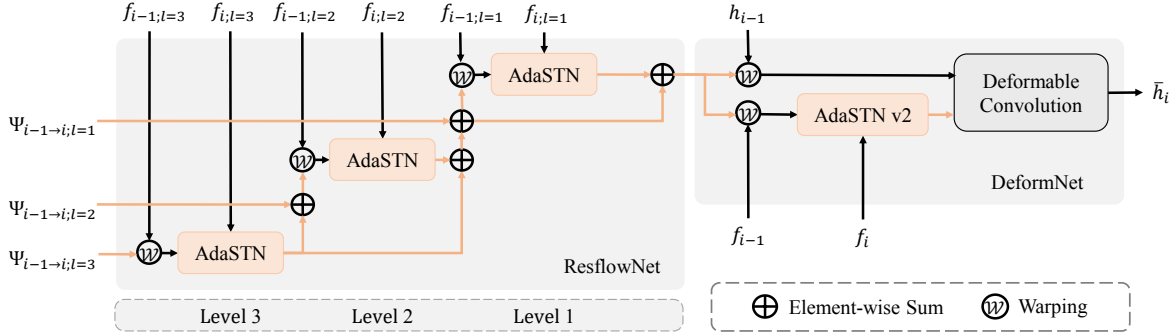
Figure 5. The structure of MultiAdaSTN. Colored lines and black lines represent the transfer direction of offsets and features, respectively.
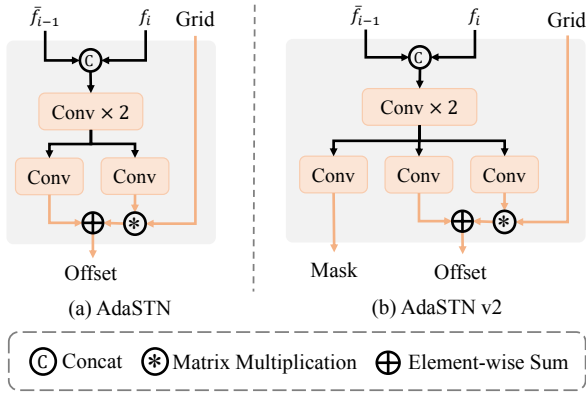

Figure 6. The structure of AdaSTN and AdaSTN v2. Colored lines and black lines represent the transfer direction of offsets and features, respectively.

We loop from Eqn. 5 to Eqn. 8 a total of l times and use the residual flow generated in the 1st-level as the ultimate residual flow $\Delta\Psi_{i-1\to i}$, *i.e.*,

$$\Delta\Psi_{i-1\to i} = \Delta\Psi_{i-1\to i;l=1}. \tag{9}$$

So far, we can obtain a more precise and task-oriented optical flow estimation. Then we present DeformNet to further refine inter-frame alignment, where modified AdaSTN (*i.e.*, AdaSTN v2) is utilized to predict the offsets and masks of deformable convolution v2 [40].

DeformNet. AdaSTN [39] predicts the pixel-level affine transformation matrix and translation vector between two misalignment features. However, it predicts the same offset for all convolutional channels and ignores the out-of-bounds issues, which reduces its flexibility. Instead, we present AdaSTN v2 to improve it, as shown in Fig. 6 (b). The predicted offset for each pixel of AdaSTN v2 can be written as,

$$\mathbf{P} = \mathbf{A}\mathbf{G} + \mathbf{b}, \tag{10}$$

where $\mathbf{A} \in \mathbb{R}^{n\times 2\times 2}$ is the estimated affine transformation matrix, $\mathbf{b} \in \mathbb{R}^{n\times 2\times 1}$ is the translation vector, and n is the

number of convolutional groups. $\mathbf{G}$ is a positional grid, which can be expressed as,

$$\mathbf{G} = \begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix}. \tag{11}$$

In addition, we add another branch to calculate masks to block out out-of-bounds and lousy information.

AdaSTN v2 can be also considered as a variant of STN [14], and it has the potential to estimate more stable offsets than deformable convolution [7]. In DeformNet, we first warp f_{i-1} and h_{i-1} to obtain aligned features by optical flow $\Psi_{i-1\to i}$, then use an AdaSTN v2 to acquire the deformable offsets $\psi_{i-1\to i}$ and masks $m_{i-1\to i}$ between these two warped features,

$$\psi_{i-1\to i}, m_{i-1\to i} = \mathcal{S}_{v2}\left(f_i, \mathcal{W}\left(f_{i-1}, \Psi_{i-1\to i}\right)\right). \tag{12}$$

Finally, we deploy a deformable convolution $\mathcal{DC}$ to get the final aligned features $\bar{h}_i$,

$$\bar{h}_i = \mathcal{DC}\left(\mathcal{W}\left(h_{i-1}, \Psi_{i-1\to i}\right), \psi_{i-1\to i}, m_{i-1\to i}\right). \tag{13}$$

4.3. Training Loss

Despite data pre-processing deployed on both our MVSR4× dataset and RealVSR [34] dataset, spatial position misalignment and color inconsistency still exist between LR-HR pairs. Learning with these data pairs will lead to pixel shifts and blurry results [38]. To alleviate the adverse effect of position and color difference, we further utilize the PWC-Net [25] and the guided filtering [9] at the patch level during training, respectively.

Denote by x_i the low-resolution frame and y_i its high-resolution counterpart, we first use guided filtering to transfer the color of y_i to close to x_i,

$$y_i^g = \mathcal{G}\left((x_i)_{\uparrow r}, y_i\right), \tag{14}$$

where $\mathcal{G}$ denotes guided filtering operation and y_i^g denotes the ground truth after converting color. Then we apply a

pre-trained optical flow network $\mathcal{F}$ (*i.e.*, PWC-Net [25]) to estimate the optical flow $\mathcal{O}$ from y_i^g to x_i,

$$\mathcal{O} = \mathcal{F}\left(x_i, (y_i^g)_{\downarrow r}\right). \tag{15}$$

The upsampled optical flow $(\mathcal{O})_{\uparrow r}$ is used to warp y_i^g,

$$y^w = \mathcal{W}\left(y_i^g, (\mathcal{O})_{\uparrow r}\right), \tag{16}$$

where y^w is a well-aligned target HR patch for supervising the VSR network. The loss for training our VSR network can be written as,

$$\mathcal{L}\left(\hat{y}_i, y^w\right) = \|\mathbf{m} \circ (\hat{y}_i - y^w)\|_1, \tag{17}$$

where $\circ$ means the pixel-wise product, and $\mathbf{m}$ is the out-of-bounds mask calculated by the optical flow $\mathcal{O}$.

5. Experiments

5.1. Implementation Details

Experiments are conducted on real-world VSR datasets RealVSR [34] and MVSR4× for training and evaluation.
Datasets. The RealVSR dataset has 450 videos for training and 50 videos for inference, and our MVSR4× randomly selects 180 videos for training and 5 videos for evaluation and leaves the remaining 15 videos for testing. The details of MVSR4× dataset can be seen in Sec. 3.
Evaluation Configurations. We conduct experiments by comparing EAVSR+ and EAVSR with 11 models: RCAN [37], FSTRN [18], TOF [33], TDAN [27], EDVR [29], RSDN [11], BasicVSR [3], ETDM [12], MANA [35], TTVSR [19], and BasicVSR++ [4]. All of these methods are retrained with ℓ_1 loss on RealVSR and MVSR4× datasets, respectively. Moreover, we train a model named EAVSRGAN+ with the perceptual loss [15] and adversarial loss [6], which is the same as the loss term of RealBasicVSR [5]. And the official model of RealBasicVSR is used for testing on RealVSR and MVSR4× datasets.

To evaluate the performance quantitatively, we compute three metrics on RGB channels, *i.e.*, Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [31], and Learned Perceptual Image Patch Similarity (LPIPS) [36], where the version of LPIPS is 0.1 trained on the AlexNet [16] network.

5.2. Experimental Results on RealVSR

As shown in Table 2, our method EAVSR+ has state-of-the-art performance between all algorithms on RealVSR [34] dataset. EAVSR only changes the alignment strategy and uses the same propagation scheme (bidirectional propagation) and reconstruction module as BasicVSR [3], and has 0.19dB PSNR gain than BasicVSR.

Besides, we show the qualitative results in Fig. 7. It can be seen that other methods cannot recover some textures on the building well. In contrast, our EAVSR+ and EAVSR can restore clearer texture details with less distortion. Furthermore, Fig. 9 (a) shows that our EAVSRGAN+ generates more pleasure results in comparison with RealBasicVSR [5]. To summarise, our methods have better performance and visual results on RealVSR dataset. Please refer to supplementary material for more qualitative results.

5.3. Experimental Results on MVSR4×

Real-world ×4 data has a much more complex degradation than ×2 data. As shown in Table. 2, on MVSR4×, our method EAVSR+ still has state-of-the-art performance between all algorithms on our MVSR4× dataset. And the gap between BasicVSR and EAVSR is increased than that in RealVSR [34] dataset. It implies that an optical flow network pre-trained on synthetic data may have performance dropping in real-world data due to the difference in degradation. In addition, EDVR [29] has a better performance than BasicVSR on ×4 real-world dataset, which is the opposite on the ×2 dataset. It also demonstrates that when degradation is more severe, the deformable convolution will be more flexible than the optical flow network, and it is not comprehensive to verify the performance of different VSR methods only on the ×2 real-world dataset.

Meanwhile, we show the qualitative comparison in Fig. 8. It can be seen that the seam of other methods between tiles is severely blurry except for EAVSR+ and EAVSR. Our EAVSR+ can recover more edges. And in Fig 9 (b), our EAVSRGAN+ can generate the result more clearly.

6. Ablation Study

In this section, we conduct extensive ablation studies on the proposed MultiAdaSTN and assess the effect of our training loss. All experiments are conducted on the RealVSR dataset [34] with EAVSR+.

We verify the performance of ResflowNet and DeformNet in MultiAdaSTN and will give a discussion about the combination of MutiAdaSTN. When neither ResflowNet nor DeformNet is applied, we directly leverage the basic offset calculated by pre-trained SPyNet [22] to warp the neighboring frames, which is considered the baseline. As shown in Table 3, benefiting from the estimation of residual offset by ResflowNet, we get 0.11 dB PSNR gain compared to the baseline. Coupled with DeformNet, the PSNR gain of 0.17 dB can be further obtained. Therefore, we concatenate these two modules together for finer results, and meanwhile will not introduce a large amount of time consumption.

In addition, We also provide the results of SPyNet and our ResflowNet on real-world data alignment in Fig. 10. The alignment result by SPyNet is warped from HR image '00028HR' by calculating the optical flow from LR im-
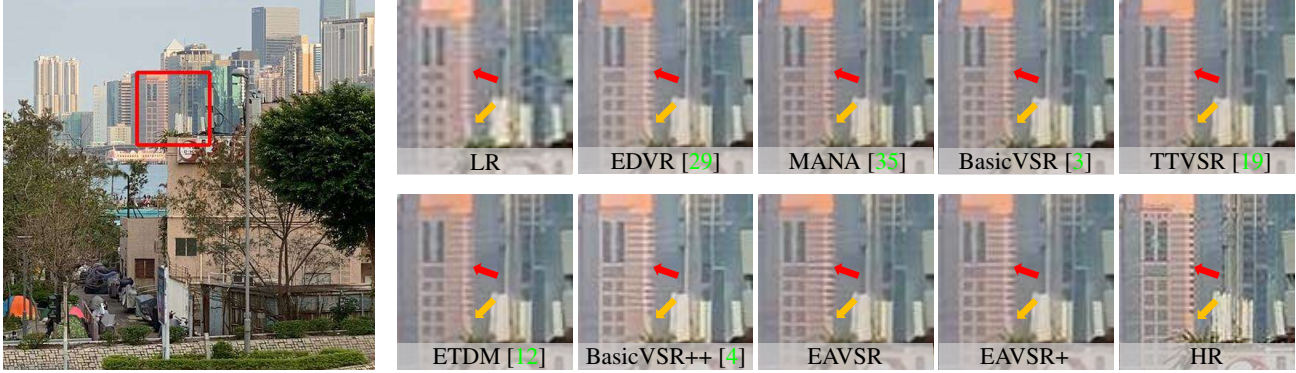
Figure 7. Visual comparison on RealVSR dataset [34]. Due to the limitation of space, here we select six methods with better PSNR. Our methods EAVSR+ and EAVSR can generate more textures, especially on the building. Please zoom in for details.
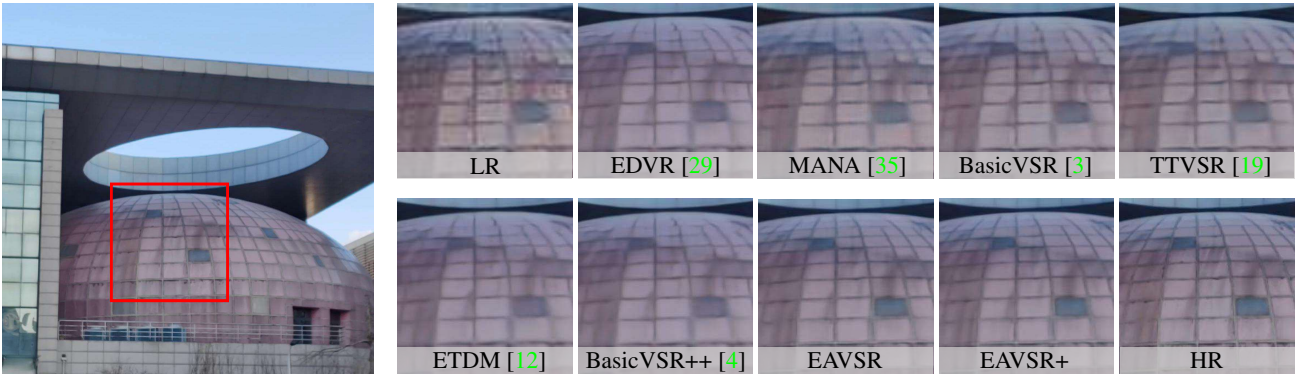

Figure 8. Visual comparison on our MVSR4× dataset. Due to the limitation of space, here we select six methods with better PSNR. Our methods EAVSR+ and EAVSR can restore sharper edges. Please zoom in for details.

Table 2. Quantitative comparison on RealVSR dataset [34] and our MVSR4× datasets. #Frame indicates the number of input frames during inference. 'p' and 'a' indicate that the previous frames and all video sequences are required, respectively. **Bold** marks the best results.

Loss Term	Method	#Frame	RealVSR [34] PSNR↑ / SSIM↑ / LPIPS↓	MVSR4× PSNR↑ / SSIM↑ / LPIPS↓
w/o Adversarial Loss	RCAN [37]	1	23.59 / 0.7720 / 0.230	22.72 / 0.7426 / 0.296
	RSDN [11]	p	23.91 / 0.7743 / 0.224	23.15 / 0.7533 / 0.279
	FSTRN [18]	7	23.36 / 0.7683 / 0.240	22.66 / 0.7433 / 0.315
	TOF [33]	7	23.62 / 0.7739 / 0.220	22.80 / 0.7502 / 0.279
	TDAN [27]	7	23.71 / 0.7737 / 0.229	23.07 / 0.7492 / 0.282
	EDVR [29]	7	23.96 / 0.7781 / 0.216	23.51 / 0.7611 / 0.268
	BasicVSR [3]	a	24.00 / 0.7801 / 0.209	23.38 / 0.7594 / 0.270
	MANA [35]	a	23.89 / 0.7781 / 0.224	23.15 / 0.7513 / 0.285
	TTVSR [19]	a	24.08 / 0.7837 / 0.213	23.60 / 0.7686 / 0.277
	ETDM [12]	a	24.13 / 0.7896 / **0.206**	23.61 / 0.7662 / 0.260
	BasicVSR++ [4]	a	24.24 / 0.7933 / 0.216	23.70 / 0.7713 / 0.263
	EAVSR (Ours)	a	24.20 / 0.7862 / 0.208	23.61 / 0.7618 / 0.264
	EAVSR+ (Ours)	a	**24.41** / **0.7953** / 0.212	**23.94** / **0.7726** / **0.259**
w/ Adversarial Loss	RealBasicVSR [5]	a	23.74 / 0.7676 / 0.174	23.15 / 0.7603 / 0.202
	EAVSRGAN+ (Ours)	a	**23.99** / **0.7726** / **0.170**	**23.35** / **0.7611** / **0.199**

(a) Visual comparison on RealVSR dataset

(b) Visual comparison on MVSR4× dataset

Figure 9. Visual comparison between RealBasicVSR [5] and our EAVSRGAN+ on RealVSR [34] and MVSR4× dataset. Please zoom in for details

00028HR   00029HR

SpyNet Alignment   Ours Alignment

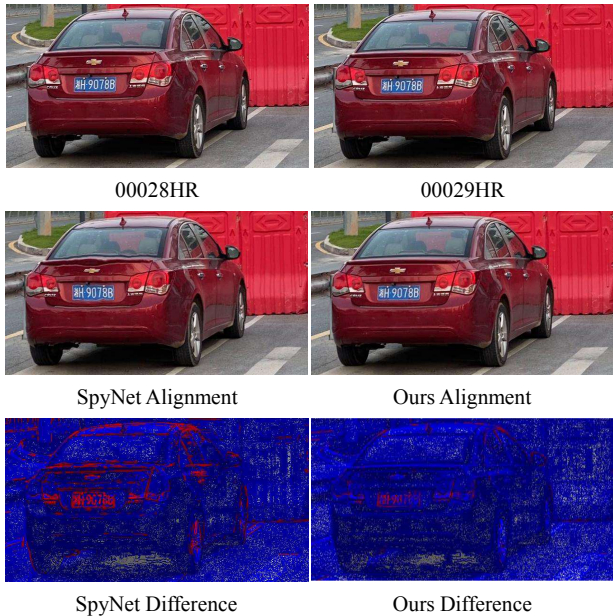SpyNet Difference   Ours Difference

Figure 10. Alignment effect comparison on SPyNet [22] and our ResflowNet. Our method can generate a more reliable offset with less distortion.

ages '00028LR' to '00029LR'. For ours, the basic offset is computed by SPyNet from '00028LR' to '00029LR' while the residual offset is computed by ResflowNet. The alignment result is warped from '00028HR' by combination of basic offset and residual offset. 'Difference' means to calculate the difference between the aligned '00028HR' image and '00029HR' image where red areas mean a large difference and blue areas mean a slight difference. Obviously, our method aligns better with introducing less distortion.

Table 3. Quantitative comparison with the different alignment combinations in MultiAdaSTN.

ResflowNet	DeformNet	PSNR	SSIM
×	×	24.13	0.7854
✓	×	24.24	0.7902
×	✓	24.20	0.7876
✓	✓	24.41	0.7953

Table 4. Quantitative comparison with the different data post-processing during training.

Position Alignment	Color Correction	PSNR	SSIM
×	×	24.32	0.7928
✓	×	24.39	0.7935
×	✓	24.34	0.7920
✓	✓	24.41	0.7953

6.1. Effect of Training Loss

6.2. Combination of MultiAdaSTN

The data pre-processing cannot completely solve the spatial position misalignment and color inconsistency issues. We leverage a robust optical flow network PWC-Net [25] to further warp the HR patch and the guided filtering [9] to convert the color of HR in the training phase. To demonstrate the advantages of the data post-processing, we verified the effect of each operation separately. Table 4 shows the results using different operations, each operation significantly improves the performance compared with the baseline. Due to space limitations, some qualitative results will show in the supplementary material.

7. Conclusion

The lack of real-world paired datasets with large scale factors and the inapplicability of alignment methods set up obstacles for existing video super-resolution (VSR) methods to apply in real scenarios. For these two issues, we build a ×4 real-world VSR dataset (named MVSR4×) and propose an effective alignment method EAVSR. The low- and high-resolution videos in the MVSR4× dataset are captured with different focal length lenses of a smartphone, respectively. EAVSR takes the proposed multi-layer adaptive spatial transform network (MultiAdaSTN) to refine the offsets provided by the optical flow estimation network. Extensive experiments show the effectiveness and practicality of our method, and we achieve state-of-the-art performance in the real-world VSR task.

8. Acknowledgement

References

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

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4(3):4, 2020. 2

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 2, 4, 6, 7

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021. 1, 2, 3, 4, 6, 7

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 1, 2, 3, 6, 7, 8

[6] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 6

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1, 2, 3, 4, 5

[8] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019. 2

[9] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 5, 8

[10] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1015–1028, 2017. 2

[11] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*, pages 645–660. Springer, 2020. 2, 6, 7

[12] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17411–17420, June 2022. 1, 3, 6, 7

[13] Takashi Isobe, Fang Zhu, Xu Jia, and Shengjin Wang. Revisiting temporal modeling for video super-resolution. *arXiv preprint arXiv:2008.05765*, 2020. 2

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 5

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[17] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17824–17833, June 2022. 3

[18] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019. 2, 3, 6, 7

[19] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5687–5696, June 2022. 1, 3, 6, 7

[20] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[21] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3043–3051, 2020. 2

[22] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 1, 2, 4, 6, 8

[23] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, 2018. 2

[24] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[25] 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, 5, 6, 8

[26] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017. 2

[27] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu Tdan. temporally-deformable alignment network for video super-resolution. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3357–3366, 2020. 2, 3, 6, 7

[28] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020. 2

[29] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3, 6, 7

[30] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 1, 2

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[32] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)*, 38(4):1–18, 2019. 2

[33] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2, 3, 6, 7

[34] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4781–4790, 2021. 1, 2, 3, 5, 6, 7, 8

[35] Jiyang Yu, Jingen Liu, Liefeng Bo, and Tao Mei. Memory-augmented non-local attention for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022. 3, 6, 7

[36] 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 conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[37] 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 the European conference on computer vision (ECCV)*, pages 286–301, 2018. 6, 7

[38] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4348–4358, October 2021. 5

[39] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, Yunjin Chen, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual zoomed observations, 2022. 2, 4, 5

[40] 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. 1, 2, 5