# Space-Time Distillation for Video Super-Resolution

Zeyu Xiao Xueyang Fu Jie Huang Zhen Cheng Zhiwei Xiong*
University of Science and Technology of China

## Abstract

*Compact video super-resolution (VSR) networks can be easily deployed on resource-limited devices, e.g., smartphones and wearable devices, but have considerable performance gaps compared with complicated VSR networks that require a large amount of computing resources. In this paper, we aim to improve the performance of compact VSR networks without changing their original architectures, through a knowledge distillation approach that transfers knowledge from a complicated VSR network to a compact one. Specifically, we propose a space-time distillation (STD) scheme to exploit both spatial and temporal knowledge in the VSR task. For space distillation, we extract spatial attention maps that hint the high-frequency video content from both networks, which are further used for transferring spatial modeling capabilities. For time distillation, we narrow the performance gap between compact models and complicated models by distilling the feature similarity of the temporal memory cells, which are encoded from the sequence of feature maps generated in the training clips using ConvLSTM. During the training process, STD can be easily incorporated into any network without changing the original network architecture. Experimental results on standard benchmarks demonstrate that, in resource-constrained situations, the proposed method notably improves the performance of existing VSR networks without increasing the inference time.*

## 1. Introduction

Video super-resolution (VSR) aims to generate a high-resolution (HR) video from its corresponding low-resolution (LR) observation. In the deep learning era, a variety of elaborately designed VSR networks can achieve promising super-resolution performance, yet at the cost of a large amount of computing resources. It is thus difficult to deploy complicated VSR networks on resource-limited devices, *e.g.*, smartphones and wearable devices. On the other hand, compact VSR networks can be easily deployed on these devices due to their lightweight architectures. However, their ability to model spatial-temporal correlations is meanwhile limited due to simple architectures, which further limits their super-resolution performance.

*Correspondence should be addressed to zwxiong@ustc.edu.cn

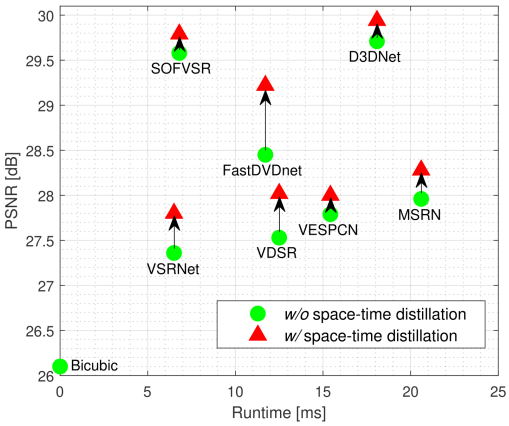



Figure 1: Comparisons on the runtime and the reconstruction quality (PSNR) of different methods (both SISR and VSR methods are included). The running time is the average execution time for super-resolving a video clip of spatial resolution $180 \times 120$ with the scale factor equal to 4 on an NVIDIA 1080Ti GPU. The PSNR value refers to the average over Vid4-Walk [41].

Different from single image super-resolution (SISR) [5, 39, 42, 40, 4, 18, 48, 2], a key step in VSR is to align different frames, either explicitly or implicitly. A majority of VSR methods contain the motion compensation module. For example, Kappeler *et al*. [17] slightly modify SRCNN [4] and extract features from frames that are aligned by optical flow. However, estimating optical flow itself is a challenging and time-consuming task [13, 31]. Inaccurate estimated optical flow leads to artifacts in the flow-based VSR methods. To avoid explicitly calculating optical flow, some recent methods exploit the motion information in an implicit manner. For example, the dynamic upsampling filters [16] and the progressive fusion residual blocks [43] are designed to explore flow-free motion compensation. It is especially worth mentioning that, as the winner of the NTIRE2019 challenges on video restoration and enhancement [24, 25], EDVR [37] utilizes a combination of pyramid, cascading, and deformable structures for multi-frame alignment. Together with temporal and spatial attention modules for information fusion, EDVR achieves state-of-the-art VSR results. Although implicit frame alignment improves computational efficiency to a certain extent, these large and complex networks still require considerable computing resources and are not competent to resource-

constrained scenarios.

To reduce the computational cost and/or required memory, a few works use recurrent schemes to introduce cost-effective network architectures for VSR [6, 14, 28]. By simply propagating the output and hidden state of previous steps with a recurrent unit, these methods achieve promising reconstruction performance and greatly reduce inference time. Although a good tradeoff between effectiveness and efficiency can be obtained, designing such recurrent networks requires tremendous efforts.

In this paper, we explore a new direction for effective and efficient VSR. Instead of pursuing more advanced network design, we introduce knowledge distillation (KD) [10] to the VSR task for the first time, which leverages intrinsic information of a teacher network to train a student network. Without changing the original architecture or increasing the inference time of the student network (a compact one), its performance is expected to be elevated towards its teacher (a complicated one). The proposed method is especially suitable for resource-limited devices, *e.g.*, smartphones and wearable devices. Once a more powerful teacher network is available, we only need to retrain the student network instead of deploying a new one.

To narrow the performance gap between compact models and complicated models, we propose a novel space-time distillation (STD) scheme to help the training of compact networks. Specifically, for spatial-related information processing, we design a space distillation (SD) scheme to utilize the spatial attention maps derived from the teacher network as the training target of the student network. This SD scheme allows a simple student network to imitate the ability of a powerful teacher network in capturing and modeling the spatial correlation. For temporal-related information processing, the powerful teacher network has a strong ability to capture temporal correlation and maintain temporal consistency. Therefore, we design a time distillation (TD) scheme to narrow the gap between the temporal memory cells of the teacher and student networks, which are encoded using a ConvLSTM from the sequence of feature maps with a sliding-window mechanism. This TD scheme not only improves the temporal consistency but also boosts the reconstruction accuracy. All these operations are only applied during training, and the network structures remain unchanged during inference. Compared to only using a reconstruction loss (*i.e.*, the Charbonnier loss [14, 15, 37]) for training, the proposed STD scheme can obtain additional performance gains from the teacher network.

Fig. 1 demonstrate that, with the proposed STD scheme and using EDVR as the teacher, notably improved VSR results can be achieved without extra runtime for a number of existing compact networks. More comprehensive experiments are conducted on two VSR benchmarks: Vid4 [21] and Vimeo90K-Test [41], where three typical compact VSR networks, *i.e.*, VESPCN [1], VSRNet [17] and FastDVDnet [34] are included for evaluation. It is verified that our proposed STD scheme improves both the reconstruction quality and the temporal consistency of VSR results while maintaining the high inference efficiency of these networks. Due to its high flexibility and generalizability, we believe the proposed STD method could greatly facilitate VSR on resource-limited devices.

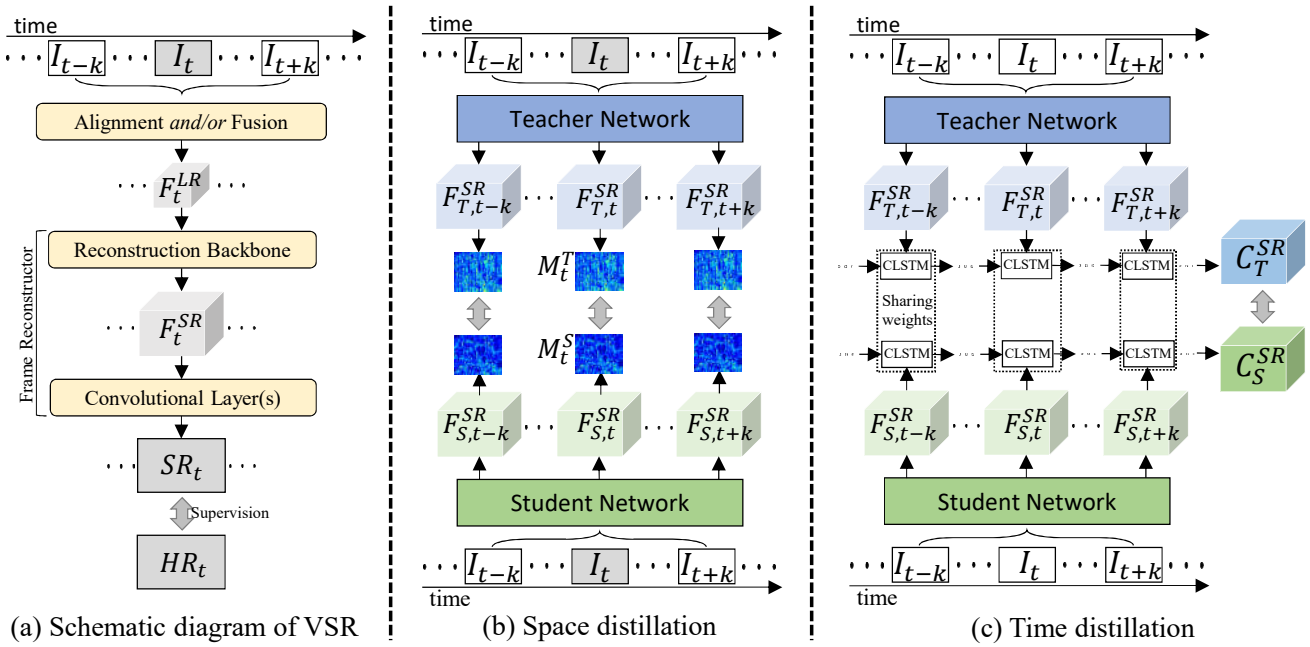
## 2. Related Work

### 2.1. Video Super-Resolution (VSR)

VSR emerges as an adaptation of SISR by exploiting additional information from neighboring low-resolution frames [6, 22, 37, 38, 41, 43, 47]. Liu *et al.* [22] propose a temporal adaptive neural network to adaptively select the optimal range of temporal dependency and a rectified optical flow alignment method for better motion estimation. Tao *et al.* [32] propose a new sub-pixel motion compensation layer for inter-frame motion alignment which can achieve motion compensation and upsampling simultaneously. Xue *et al.* [41] train motion estimation and VSR jointly in an end-to-end manner through the proposed task-oriented flow. Instead of image level motion alignment, TDAN [35] and EDVR [37] work at the feature level. TDAN [35] uses a temporal deformable alignment module to align different frames in the feature domain for better reconstruction performance. EDVR [37] further extends TDAN by using deformable alignment in a coarse-to-fine manner and a new temporal and spatial attention fusion module, instead of naively concatenating the aligned LR frames. Recently, Isobe *et al.* [14] propose a recurrent-based network in which the structure and detail components extracted from LR inputs are reconstructed with two-stream structure-detail blocks for efficient VSR.

### 2.2. Knowledge Distillation (KD)

KD refers the technique that leverages intrinsic information of a large teacher network to train a small student one. KD is first proposed by Hinton *et al.* [10], and since then many works have been devoted to this topic [9, 12, 20, 23, 27, 44, 49]. Very recently, KD has been extended to SISR and proven its effectiveness. Gao *et al.* [7] attempt to propagate the first-order statistical information (*e.g.*, average pooling over channels) from a teacher model, where a student model is trained to have similar feature distributions to that of the teacher. Lee *et al.* [19] take the ground truth HR images as inputs to extract powerful privileged information for image reconstruction. These methods can be seen as the pioneering works along the SISR line. However, KD for the VSR task remains unexplored, which is the focus of this paper.




(a) Schematic diagram of VSR (b) Space distillation (c) Time distillation

Figure 2: Our distillation framework. (a) The toy architecture of the complicated VSR method (*e.g.*, EDVR [37]). The output super-resolved frame $SR_t$ is generated by adding the residual map produced by the network and the bicubic upsampling result of the input reference frame. We have omitted this process in the figure. (b) The space distillation allows a network to exploit spatial attention maps derived from the teacher network (*e.g.*, $M_t^{\mathtt{T}}$) as the distillation targets for the spatial attention maps extracted from the student network (*e.g.*, $M_t^{\mathtt{S}}$). (c) The time distillation encodes the extracted multi-frame feature maps into temporal memory cells $C_{\mathtt{T}}^{SR}$ and $C_{\mathtt{S}}^{SR}$ by using ConvLSTM and minimizes the gap between the teacher and student networks.

## 3. Space-Time Distillation

Fig. 2(a) shows the toy architecture of a complicated VSR network. Given $2k+1$ consecutive LR frames $I_{[t-k:t+k]}$, we denote the middle frame $I_t$ as the reference frame and the other frames as neighboring frames. The aim of VSR is to estimate an upsampled reference frame $SR_t$, which is desired to be close to the ground truth $HR_t$. Each input neighboring frame is aligned to the reference frame at the feature level and/or aggregated in the spatial-temporal domain with an elaborately designed motion alignment module. We use $F_t^{LR}$ to represent the aligned and/or fused feature, and it is further sent to the reconstruction backbone with the pixel-shuffle operation [29] to estimate $F_t^{SR}$, which has the same spatial resolution as $HR_t$. The restored $SR_t$ is finally obtained by convolving $F_t^{SR}$ to reduce the number of channels.

Different from Fig. 2(a), we apply our newly designed STD scheme, as in Figs. 2(b) and 2(c), to transfer the knowledge from the complicated teacher network T to a compact student network S. Transferring such multi-frame alignment and spatial-temporal fusion capabilities from T to S could make the student mimic the teacher better in terms of video reconstruction. Note that, we choose to distill $F_t^{SR}$ instead of $F_t^{LR}$, since distilling $F_t^{SR}$ can achieve better reconstruction accuracy (as shown in Sec. 4.3).

### 3.1. Space Distillation (SD)

High-frequency details are critical to the reconstruction of reference frames. Inspired by the activation-based attention distillation [46], we design an SD scheme to model the spatial representation ability of T by extracting the spatial attention map from T, and utilize it to train the compact S. We use $F_{\mathtt{T},t}^{SR} \in \mathbb{R}^{C\times W\times H}$ and $F_{\mathtt{S},t}^{SR} \in \mathbb{R}^{C\times W\times H}$ to denote the feature maps of the teacher and student networks. $C$, $H$ and $W$ denote the channel, height and width of feature maps, respectively. The generation of the spatial attention map is equivalent to finding a mapping function $\mathcal{M}: \mathbb{R}^{C\times W\times H} \rightarrow \mathbb{R}^{W\times H}$. The spatial attention map contains diverse and rich contextual information that hints the high-frequency video content. The mapping function can be defined as one of the following three operations [46]

$$\mathcal{M}_{sum}(F_t^{SR}) = \sum_{i=1}^{C} |F_{t,i}^{SR}|, \tag{1}$$

$$\mathcal{M}_{sum}^2(F_t^{SR}) = \sum_{i=1}^{C} |F_{t,i}^{SR}|^2, \tag{2}$$

$$\mathcal{M}_{max}^2(F_t^{SR}) = max_{i=1}^{C} |F_{t,i}^{SR}|^2, \tag{3}$$

where $F_{t,i}^{SR}$ is the $i$-th slice in the channel dimension. Note that we describe the mapping function here in a unified way.

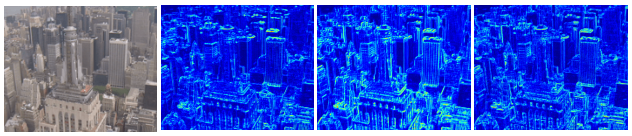



Figure 3: The input reference frame and visualization of extracted spatial attention maps using three mapping functions. The spatial attention map extracted by $\mathcal{M}^2_{sum}(\cdot)$ hints the high-frequency video content more accurately.

Both student and teacher can be represented by adding S and T subscripts, respectively.

We visualize the spatial attention maps of these three mapping functions in Fig. 3. Compared with $\mathcal{M}_{sum}(F_t^{SR})$, $\mathcal{M}^2_{sum}(F_t^{SR})$ assigns more weights to areas with high-frequency details. Compared with $\mathcal{M}^2_{max}(F_t^{SR})$, $\mathcal{M}^2_{sum}(F_t^{SR})$ describes the details of the scene more clearly and accurately since it calculates weights in a global mechanism rather than simply selecting the maximum value. Based on the above visualization and analysis, in the following experiment, we use $\mathcal{M}^2_{sum}(F_t^{SR})$ as the mapping function which yields the best performance.

By using the mapping function $\mathcal{M}^2_{sum}(\cdot)$, the *spatial attention maps* of the teacher and student networks can be calculated as

$$M_t^{\mathrm{T/S}} = \mathcal{M}^2_{sum}(F^{SR}_{\mathrm{T/S},t}). \tag{4}$$

During training the student network, the spatial attention map $M_t^{\mathrm{S}}$ is forced to be close to $M_t^{\mathrm{T}}$. Transferring knowledge contained in spatial attention maps from the teacher to the student could make the student mimic the teacher better in terms of learning high-frequency details, and thus improve the performance of the student network. The SD loss to optimize the student S is

$$\mathcal{L}_{SD} = \frac{1}{N}\sum_{t=1}^{N}\mathcal{L}_d(M_t^{\mathrm{S}}, M_t^{\mathrm{T}}), \tag{5}$$

where $\mathcal{L}_d$ is typically defined as the $L_2$-norm distance and $N$ is the number of frames in a training clip. We use the sliding-window scheme to create training pairs. For the boundary frames, we create pairs by duplicating these frames for multiple times.

### 3.2. Time Distillation (TD)

Exploiting the correlation among multiple frames is the key step in VSR. The complicated teacher network has a stronger ability to handle temporal information with large motions due to its well-designed frame alignment and/or fusion structures. Our TD scheme, as shown in Fig. 2(c), is designed to migrate the temporal modeling ability of the teacher network to the student network.

Given $2k+1$ consecutive LR frames $I_{[t-k:t+k]}$, the corresponding feature maps $F^{SR}_{[t-k:t+k]}$ are extracted from the output of the reconstruction backbone before the convolutional layer(s). We input these $2k+1$ feature maps into ConvLSTM units [30] in order. By continuously passing the hidden states of the previous features to the last units, the output $C^{SR}$, called temporal memory cell, can record the long-term temporal information of the input video clip. We use $C^{SR}_{\mathrm{T}}$ and $C^{SR}_{\mathrm{S}}$ to denote the temporal memory cells of the teacher and student networks encoded by ConvLSTM units as

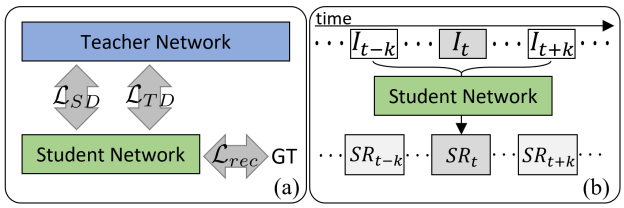



Figure 4: Our training and inference mechanisms. (a) The training mechanism. The space distillation loss, the time distillation loss and the reconstruction loss are used during training. (b) The inference mechanism. The proposed SD and TD schemes are only applied during training, and the network structures remain unchanged during inference.

$$(C^{SR}_{\mathrm{T/S}}, h_{\mathrm{T/S},t+k}) = ConvLSTM(F^{SR}_{\mathrm{T/S},t+k-1}, h_{\mathrm{T/S},t+k-1}), \tag{6}$$

where $h_{\mathrm{T/S},t+k}$ represents the hidden states of the teacher network T or the student network S at time $t+k$. The complete definition of $ConvLSTM(\cdot)$ are specified in the supplementary document due to space limitation.

The proposed TD scheme aims to minimize the gap between temporal memory cells $C^{SR}_{\mathrm{T}}$ and $C^{SR}_{\mathrm{S}}$. The TD loss used to optimize the student S is

$$\mathcal{L}_{TD} = \mathcal{L}_d(C^{SR}_{\mathrm{T}}, C^{SR}_{\mathrm{S}}), \tag{7}$$

where $\mathcal{L}_d$ is typically defined as the $L_2$-norm distance. The network parameters in the ConvLSTM units are optimized together with the student network. To extract the multi-frame temporal information, both the teacher and student networks share the weights of the ConvLSTM units. Note that there may exist a model collapse point when the weights and biases in the ConvLSTM units are all equal to zero. During training, when the value of the TD loss is lower than $1e^{-8}$, we fix the parameters of ConvLSTM to prevent the model collapse.

### 3.3. Loss Functions

We use the Charbonnier loss [37, 14, 15] as the reconstruction loss function to further constrain the reconstructed results

$$\mathcal{L}_{rec} = \sqrt{\|SR_t - HR_t\|^2 + \varepsilon^2}, \tag{8}$$

where $\varepsilon$ is set to $1e^{-6}$. The complete loss function for training a compact student network S is

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{SD} + \lambda_2 \mathcal{L}_{TD}, \tag{9}$$

where $\lambda_1$ and $\lambda_2$ are weighting factors. The proposed SD (Eq. 5) and TD (Eq. 7) schemes are only applied during training, together with the reconstruction loss as shown in Fig. 4(a). During inference, the network structure remains unchanged as shown in Fig. 4(b).

# 4. Experiments

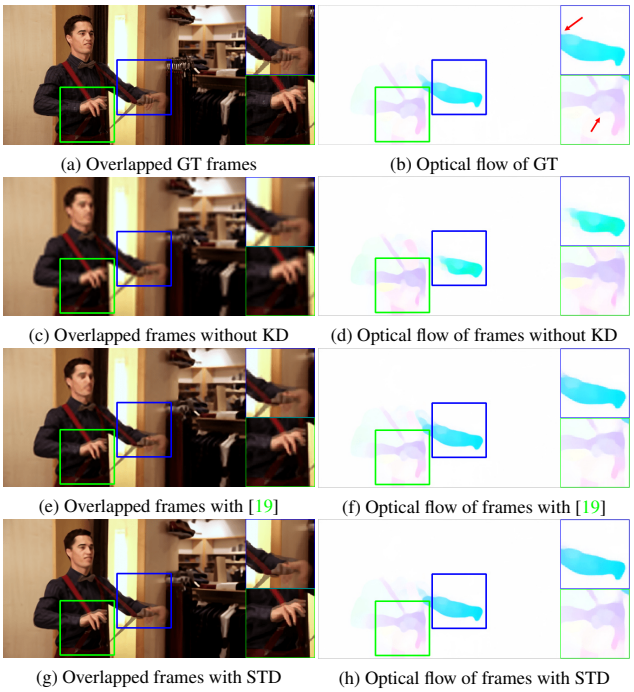
## 4.1. Experimental Settings

**Network Structures**. We adopt the state-of-the-art VSR network EDVR [37] as the complicated teacher network T and employ several simpler and shallower networks as students to verify the effectiveness of our STD scheme. We first consider FastDVDnet [34] as a basic student network and conduct ablation studies on it. Since FastDVDnet [34] is originally proposed for video denoising [3, 33, 34], we change its structure to make it suitable for the VSR task (more details can be found in the supplementary document). Typical compact VSR networks like VSRNet [17], VESPCN [1], D3DNet [45] and SOFVSR [36] are also adopted in our experiments.

**Training Settings**. We follow the same experimental setting as in [14] and train our models on the Vimeo90K [41] dataset. We crop patches of size $256 \times 256$ from HR video clips as the target. The corresponding LR patches are obtained by applying Gaussian blur ($\sigma = 1.6$) to the target patches followed by $4\times$ times downsampling. Rotation and flipping are applied for data augmentation.

**Inference Settings**. We evaluate our STD scheme on several popular benchmarks, including Vimeo90K-Test [41] (excluding the training data) and Vid4 testset [21]. Vimeo90K-Test contains about 8K high-quality clips with diverse motion types (*i.e.*, fast, medium and slow motion). Vid4 consists of four scenes with various motion and occlusion, which is useful to evaluate the robustness of different methods. To quantitatively evaluate the reconstructed video, we choose PSNR and SSIM [11] as the main metrics. Temporal profile [16, 14, 38, 50] is also included to evaluate the temporal consistency qualitatively.

**Implementation Details**. We utilize the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each mini-batch consists of 2 samples. The learning rate is initially set to $1e^{-4}$ and is later down-scaled by a factor of 0.5 every 200K iterations till 800K iterations. We set $\lambda_1 = 1$, $\lambda_2 = 100$ and $k = 3$. We use $\mathcal{L}_{rec}$ (Eq. 8) to warm up the student networks for 20K iterations, and then use $\mathcal{L}$ (Eq. 9) to complete the remaining training. Experiments are conducted using PyTorch [26] on NVIDIA 1080Ti GPUs.



(a) Overlapped GT frames (b) Optical flow of GT
(c) Overlapped frames without KD (d) Optical flow of frames without KD
(e) Overlapped frames with [19] (f) Optical flow of frames with [19]
(g) Overlapped frames with STD (h) Optical flow of frames with STD

Figure 5: Visual comparisons on Vimeo90K-Test. We use the FlowNet2.0 [13] to estimate optical flow of two adjacent frames generated by VDSR [18]. The optical flow of the distilled model with our STD scheme has smoother shapes and is closer to GT. Please zoom in for better visualization.

## 4.2. Quantitative and Qualitative Comparisons

We evaluate the distilled student models against their baselines which are trained with the Charbonnier loss in terms of PSNR/SSIM. Other VSR methods, including Bicubic, TOFlow [41], VSR-DUF [16], RBPN [8], EDVR [37] are also included for comparison. We also report the FLOPs (TMAC) required to reconstruct the frame with the spatial resolution of $180 \times 120$. The average runtime of each method measured on Vid4-Walk in a per-frame manner is presented using one 1080Ti GPU. Note that we also compare the results of the distillation scheme in [19] and our proposed STD scheme.

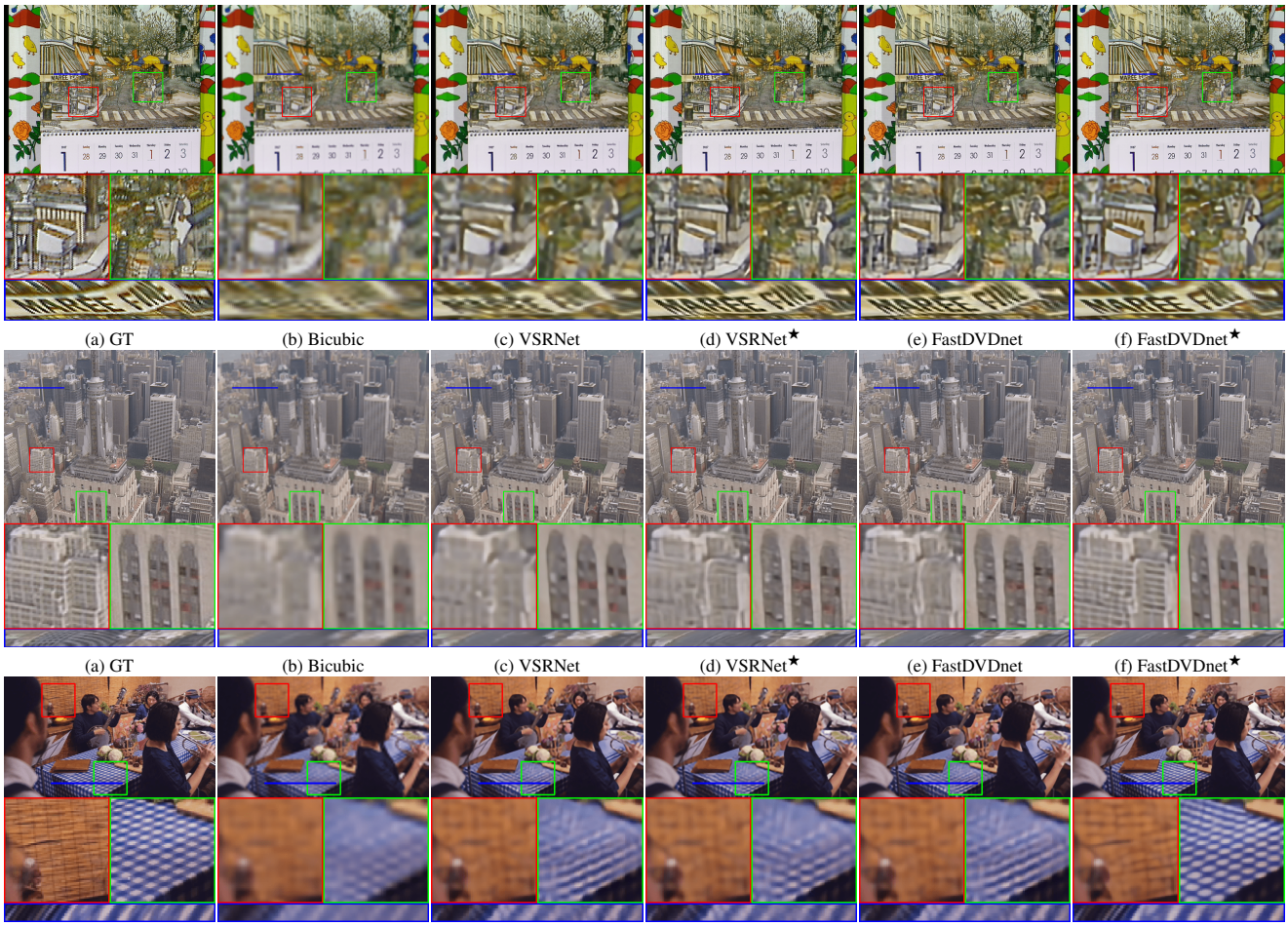
Table 1 shows quantitative comparisons on Vid4 and Vimeo90K-Test. From the table, we can observe that:

**(1)** The student models trained with STD outperform the corresponding baselines with a considerable margin. Specifically, on Vid4, FastDVDnet trained with STD achieves 26.14dB (PSNR), while the same model trained without STD only gets 25.40dB (PSNR).

**(2)** The student networks trained with STD achieve higher reconstruction performance compared with the distillation scheme in [19]. We visualize the superposition results and optical flow of two reconstructed adjacent frames in Fig. 5. It can be observed that using our STD scheme

Table 1: Quantitative comparisons of different methods on Vid4 and Vimeo90K-Test for $4\times$ upscaling in terms of PSNR (dB). Results are evaluated on the Y (luminance) channel. 'Frames' means the number of input frames of the network. 'FLOPs' (T, $10^{12}$) is calculated on a frame with the spatial resolution of $180 \times 120$. 'Time' is the average running time (ms) which is measured on Vid4-Walk in a per-frame manner. ★ means the student network is trained with our STD scheme and ♣ means the student network is trained with the scheme proposed in [19].

| Method | Frames | Network performance | | Vid4 | | | | | Vimeo90K-Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FLOPs | Time | Calendar | City | Foliage | Walk | Average | Fast | Medium | Slow | Average |
| Bicubic | 1 | - | - | 20.39 | 25.16 | 23.47 | 26.10 | 23.78 | 32.99 | 30.24 | 28.28 | 30.28 |
| TOFlow | 7 | 0.81 | 632.0 | 22.29 | 26.79 | 25.31 | 29.02 | 25.84 | 37.64 | 35.02 | 32.16 | 34.84 |
| VSR-DUF | 7 | 0.62 | 496.0 | 24.04 | 28.05 | 26.41 | 30.60 | 27.33 | 37.49 | 35.84 | 32.96 | 35.50 |
| EDVR | 7 | 0.93 | 86.0 | 23.75 | 28.27 | 26.03 | 30.51 | 27.14 | 39.67 | 37.18 | 34.14 | 36.89 |
| VDSR | 1 | 0.22 | 11.5 | 21.09 | 25.66 | 24.14 | 27.53 | 24.61 | 35.92 | 33.04 | 30.80 | 33.04 |
| VDSR♣ | 1 | 0.22 | 11.5 | 21.22 | 25.90 | 24.31 | 27.66 | 24.77 | 36.34 | 33.32 | 31.14 | 33.34 |
| VDSR★ | 1 | 0.22 | 11.5 | 21.55 | 25.84 | 24.33 | 28.02 | 24.94 | 36.79 | 33.90 | 31.58 | 33.89 |
| VESPCN | 3 | 0.26 | 21.3 | 21.16 | 25.91 | 24.51 | 27.78 | 24.84 | 36.22 | 33.38 | 31.12 | 33.36 |
| VESPCN★ | 3 | 0.26 | 21.3 | 21.32 | 26.25 | 24.81 | 27.99 | 25.09 | 36.64 | 33.89 | 31.57 | 33.84 |
| VSRNet | 7 | 0.23 | 11.3 | 21.00 | 25.60 | 24.14 | 27.36 | 24.52 | 36.09 | 33.27 | 31.00 | 33.24 |
| VSRNet★ | 7 | 0.23 | 11.3 | 21.33 | 25.74 | 24.33 | 27.80 | 24.80 | 36.52 | 33.76 | 31.45 | 33.72 |
| FastDVDnet | 7 | 0.06 | 17.5 | 22.11 | 26.34 | 24.80 | 28.45 | 25.43 | 37.25 | 34.54 | 32.11 | 34.48 |
| FastDVDnet★ | 7 | 0.06 | 17.5 | 22.71 | 27.18 | 25.46 | 29.22 | 26.14 | 38.13 | 36.07 | 33.68 | 36.12 |

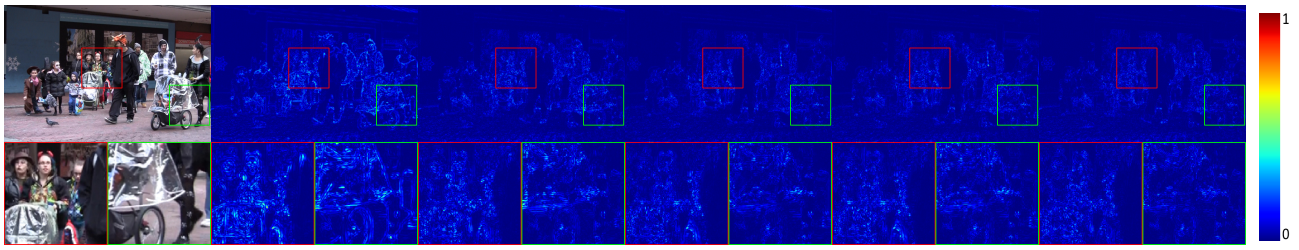


(a) GT (b) Bicubic (c) VSRNet (d) VSRNet★ (e) FastDVDnet (f) FastDVDnet★

(a) GT (b) Bicubic (c) VSRNet (d) VSRNet★ (e) FastDVDnet (f) FastDVDnet★

(a) GT (b) Bicubic (c) VSRNet (d) VSRNet★ (e) FastDVDnet (f) FastDVDnet★

Figure 6: Visual comparisons of the distilled student models against their baselines on $4\times$ upscaling. ★ means student networks trained with our STD scheme. The reconstructed frames, and the temporal profiles at the blue scan lines are provided. The frames are from Vid4-Calendar, Vid4-City and Vimeo90K-Test-00001/0629, respectively. Please zoom in for better visualization.




(a) GT (b) Bicubic (c) FastDVDnet -*w/o* STD (d) FastDVDnet -*w/* SD (e) FastDVDnet -*w/* TD (f) FastDVDnet -*w/* STD

Figure 7: Error map comparison on the frame from Vid4-Walk. Please zoom in for better visualization.

obtains results with better temporal consistency.

**(3)** The models trained with STD offer a good compromise in terms of PSNR and the number of operations/runtime. For example, TOFlow [41] has 0.81T operations and average runtime of 632ms to achieve the average PSNR of 25.84dB on Vid4. On the contrary, our STD scheme further boosts FastDVDnet without modifying the network architecture, achieving the average PSNR of 26.14dB and has 0.06T operations only, while taking 17.5ms for inference.

In Fig. 1, we provide more comparisons on the effectiveness of our proposed STD scheme. As can be seen, we can achieve better reconstruction results with existing compact networks, but without extra runtime.

Qualitative results on Vid4 are presented in Fig. 6. It is clear that the student models trained with STD provide better qualitative results than their baselines with more accurate details and less blurs. As can be seen from the accompanied temporal profiles, methods trained with STD are able to reconstruct temporally consistent results while their baselines incur obvious temporal inconsistency.

### 4.3. Model Analysis

To analyze the flexibility and generalizability of the proposed STD scheme, we conduct the following experiments on the Y channel with the metric of PSNR.

**Effectiveness of space-time distillation**. We first investigate the contribution of SD and TD by taking FastDVDnet as an example. We show the quantitative results in Table 2, and provide the error map comparison in Fig. 7. We observe that the model trained with only SD (-*w/* SD) or TD (-*w/* TD) achieves higher PSNR than the baseline model (-*w/o* STD), while the model trained with both SD and TD (-*w/* STD) achieves the best results. Specifically, SD and TD provide about 0.49dB and 0.21dB PSNR gain on Vid4 compared with the baseline. Using SD and TD simultaneously provides 0.74dB gain in terms of PSNR compared with the baseline. This shows that STD can boost the performance of the student network by effectively transferring spatial-temporal information from the teacher. As shown in Fig. 7, the result of the model trained with STD has less errors compared with the baseline.

Table 2: Analysis on the effectiveness of the proposed STD scheme. Experiments are conducted on Vid4.

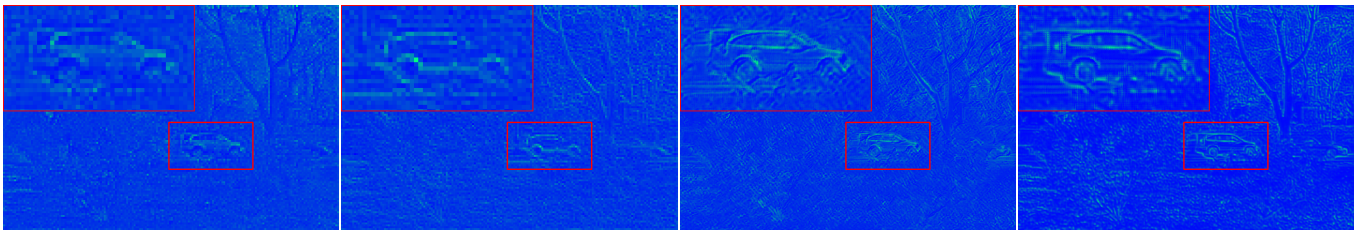
| Vid | Calendar | City | Foliage | Walk | Average |
|---|---|---|---|---|---|
| FastDVDnet -*w/o* STD | 22.11 | 26.34 | 24.80 | 28.45 | 25.43 |
| FastDVDnet -*w/* SD | 22.63 | 27.09 | 25.19 | 28.90 | 25.95 |
| FastDVDnet -*w/* TD | 22.59 | 26.51 | 25.26 | 29.14 | 25.87 |
| FastDVDnet -*w/* STD | 22.71 | 27.18 | 25.46 | 29.22 | 26.14 |

Table 3: Analysis on different distillation schemes. Experiments are conducted on Vid4 using FastDVDnet.

| Vid4 | Calendar | City | Foliage | Walk | Average |
|---|---|---|---|---|---|
| KD - $Ours$ | 22.71 | 27.18 | 25.46 | 29.22 | 26.14 |
| KD - $MSE$ | 22.56 | 26.89 | 25.14 | 28.83 | 25.86 |
| KD-$\mathcal{M}^2_{mean}$ | 22.61 | 26.82 | 25.23 | 29.02 | 25.92 |

**Our space-time distillation scheme vs. other distillation schemes**. We compare our STD with other two types of distillation schemes using FastDVDnet. The first one is the simplest distillation scheme which uses MSE loss to constrain the feature similarity between the teacher and student networks (KD-$MSE$). The other one is from [7], denoted as KD-$\mathcal{M}^2_{mean}$, in which the feature gap is narrowed by $\mathcal{M}^2_{mean}(F^{LR}_t) = (\frac{1}{C}\sum_{i=1}^{C} F^{LR}_{t,i})^2$. The quantitative results displayed in Table 3 show that our STD performs better than these two distillation schemes. STD provides an average of 0.28dB PSNR gain on Vid4 over KD-$MSE$. Compared with the student model trained with KD-$\mathcal{M}^2_{mean}$, the student trained with our STD scheme achieves a 0.22dB increase in PSNR. This demonstrates that our STD can better transfer the teacher's ability of modeling spatial-temporal information to the student.

**Effectiveness of distilling features with higher resolution**. In experiment, we empirically find that using the high-resolution feature $F^{SR}_t$ for distillation is more effective than using the low-resolution feature $F^{LR}_t$. Here we show the quantitative results on distilling $F^{LR}_t$ (KD-$F^{LR}_t$) and $F^{SR}_t$ (KD-$F^{SR}_t$) of FastDVDnet in Table 4. Compared to using $F^{LR}_t$ for distillation, distilling $F^{SR}_t$ achieves an average increase of 0.05dB in terms of PSNR on Vid4. We also visualize the representative feature maps of EDVR-$F^{LR}$, EDVR-$F^{SR}$, KD-$F^{LR}$ and KD-$F^{LR}$ in the same channel. Feature visualization in Fig. 8 validates this observation. $F^{SR}_t$ con-



(a) EDVR-$F_t^{LR}$ (b) KD-$F_t^{LR}$ (c) EDVR-$F_t^{SR}$ (d) KD-$F_t^{SR}$

Figure 8: Feature maps ($F_t^{LR}$ and $F_t^{SR}$) extracted from EDVR and FastDVDnet after distillation on Vid4-Foliage.

Table 4: Analysis on distilling features with different resolutions. Experiments are conducted on Vid4 using FastDVDnet.

| Vid4 | Calendar | City | Foliage | Walk | Average |
|---|---|---|---|---|---|
| KD - $F_t^{SR}$ | 22.71 | 27.18 | 25.46 | 29.22 | 26.14 |
| KD - $F_t^{LR}$ | 22.66 | 27.12 | 25.43 | 29.14 | 26.09 |

Table 5: Using RBPN (-RBPN) and EDVR (-EDVR) to distill FastDVDnet based on our proposed STD scheme. Experiments are conducted on Vid4.

| Vid4 | Calendar | City | Foliage | Walk | Average |
|---|---|---|---|---|---|
| EDVR | 23.75 | 28.27 | 26.03 | 30.51 | 27.14 |
| RBPN | 23.70 | 27.59 | 25.94 | 30.36 | 26.90 |
| FastDVDnet | 22.11 | 26.34 | 24.80 | 28.45 | 25.43 |
| FastDVDnet-RBPN | 22.45 | 26.84 | 25.33 | 28.96 | 25.89 |
| FastDVDnet-EDVR | 22.71 | 27.18 | 25.46 | 29.22 | 26.14 |

tains more details, which can be better transferred from the teacher network to the student network.

**Using different teachers for distillation**. Can the student network be able to learn from different teachers based on our proposed STD scheme? To answer this question, here we use another teacher network, *i.e.*, RBPN [8], a typical optical-flow-based VSR method, to verify the versatility of STD, and the results are shown in Table 5. We make minor adjustments to the last convolutional layer of RBPN to make it suitable for our STD scheme. Using RBPN as the teacher can also obtain performance improvement: compared to FastDVDnet without distillation, distilling the same network using RBPN has a 0.46dB increase in PSNR. Although there is a certain gap with the distilled results using EDVR as the teacher, we can draw the following conclusions. First, our STD scheme is applicable to different teachers and students. Second, the better the teacher's performance, the larger the distillation improvement. If a better network than EDVR is adpoted as the teacher for distillation, larger performance improvement can be achieved by using our STD scheme. Due to its high flexibility and generalizability, we believe the proposed STD scheme could greatly facilitate VSR on resource-limited devices.

**Distilling the student network with different numbers of parameters**. To verify the robustness of STD to models of different sizes, *i.e.*, models with different numbers of parameters, we use STD to distill FastDVDnet as an example. Specifically, we change the number of feature channels in front of the reconstruction backbone in FastDVDnet to $1/2$ and $1/4$ of the original one (*i.e.*, the numbers of channels in FastDVDnet are set to 32 and 16, respectively). Table 6 shows the results of FastDVDnet without STD (*-w/o* STD) and with STD (*-w/* STD) under different model sizes. We find that the STD scheme can improve the network performance under different model sizes, especially when the model size is small. For example, when the model size of FastDVDnet is $1/2$ of the original one, the model trained with STD achieves 0.68dB gain over that without STD; when the model size shrinks to only 1/4 of the original one, STD provides 1.36dB gain. This is reasonable: when the model size is small, its ability to model spatial-temporal correlations is limited. Therefore, the knowledge transferred from the teacher network promotes the performance of the student network more.

Table 6: The reconstruction quality (PSNR) of FastDVDnet trained with and without the proposed STD scheme under different model sizes on Vid4-Walk.

| Model size | 1/4 | 1/2 | 1 | Bicubic |
|---|---|---|---|---|
| FastDVDnet *-w/o* STD | 25.70 | 26.67 | 28.45 | 26.10 |
| FastDVDnet *-w/* STD | 27.06 | 27.35 | 29.22 | |

## 5. Conclusion

In this work, we propose a novel knowledge distillation scheme, *i.e.*, space-time distillation, for VSR in resource-constrained situations. Our method is able to train compact student networks with the help of complicated teacher networks. We demonstrate the effectiveness, flexibility and versatility of our proposed distillation scheme on the VSR task, which can achieve higher reconstruction quality using existing VSR methods, while maintaining their small model sizes and fast inference time. In future work, we will explore more suitable distillation losses for VSR to further improve the performance of compact networks.

## Acknowledgement

We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, National Natural Science Foundation of China under Grant 61901433, and the USTC Research Funds of the Double First-Class Initiative under Grant YD2100002003.

## References


[1] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017. 2, 5

[2] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, 2019. 1

[3] Michele Claus and Jan van Gemert. Videnn: Deep blind video denoising. In *CVPR Workshops*, 2019. 5

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1

[5] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1):25–47, 2000. 1

[6] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCV Workshops*, 2019. 2

[7] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. Image super-resolution using knowledge distillation. In *ACCV*, 2018. 2, 7

[8] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019. 5, 8

[9] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 2

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS*, 2015. 2

[11] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010. 5

[12] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *ICCV*, 2019. 2

[13] 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. 1, 5

[14] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, 2020. 2, 4, 5

[15] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *CVPR*, 2020. 2, 4

[16] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018. 1, 5

[17] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016. 1, 2, 5

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 1, 5

[19] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *ECCV*, 2020. 2, 5, 6

[20] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, 2017. 2

[21] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):346–360, 2013. 2, 5

[22] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, 2017. 2

[23] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. Efficient semantic video segmentation with per-frame inference. In *ECCV*, 2020. 2

[24] Seungjun Nah, Radu Timofte, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring: Methods and results. In *CVPR Workshops*, 2019. 1

[25] Seungjun Nah, Radu Timofte, Shuhang Gu, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, and Kyoung Mu Lee. Ntire 2019 challenge on video super-resolution: Methods and results. In *CVPR Workshops*, 2019. 1

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 2

[28] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 2

[29] 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 *CVPR*, 2016. 3

[30] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 4

[31] 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. 1

[32] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 2

[33] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *ICIP*, 2019. 5

[34] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *CVPR*, 2020. 2, 5

[35] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2

[36] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through hr optical flow estimation. In *ACCV*, 2018. 5

[37] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019. 1, 2, 3, 4, 5

[38] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. Space-time video super-resolution using temporal profiles. In *ACM MM*, 2020. 2, 5

[39] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Image hallucination with feature enhancement. In *CVPR*, 2009. 1

[40] Zhiwei Xiong, Dong Xu, Xiaoyan Sun, and Feng Wu. Example-based super-resolution with soft information and decision. *IEEE Transactions on Multimedia*, 15(6):1458–1465, 2013. 1

[41] 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. 1, 2, 5, 7

[42] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010. 1

[43] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 1, 2

[44] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 2

[45] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters*, 27, 2020. 5

[46] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3

[47] Haochen Zhang, Dong Liu, and Zhiwei Xiong. Two-stream action recognition-oriented video super-resolution. In *ICCV*, 2019. 2

[48] 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. 1

[49] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *CVPR*, 2019. 2

[50] Liad Pollak Zuckerman, Shai Bagon, Eyal Naor, George Pisha, and Michal Irani. Across scales & across dimensions: Temporal super-resolution using deep internal learning. In *ECCV*, 2020. 5