Ada-VSR: Adaptive Video Super-Resolution with Meta-Learning

Akash Gupta, Padmaja Jonnalagedda, Bir Bhanu, Amit K. Roy-Chowdhury
University of California, Riverside
{agupt013@, sjonn002@, bir.bhanu@, amitrc@ece}.ucr.edu

ABSTRACT

Most of the existing works in supervised spatio-temporal video super-resolution (STVSR) heavily rely on a large-scale external dataset consisting of paired low-resolution low-frame rate (LR-LFR) and high-resolution high-frame rate (HR-HFR) videos. Despite their remarkable performance, these methods make a prior assumption that the low-resolution video is obtained by down-scaling the high-resolution video using a known degradation kernel, which does not hold in practical settings. Another problem with these methods is that they cannot exploit instance-specific internal information of a video at testing time. Recently, deep internal learning approaches have gained attention due to their ability to utilize the instance-specific statistics of a video. However, these methods have a large inference time as they require thousands of gradient updates to learn the intrinsic structure of the data. In this work, we present **Ada**ptive **V**ideo **S**uper-**R**esolution (**Ada-VSR**) which leverages external, as well as internal, information through meta-transfer learning and internal learning, respectively. Specifically, meta-learning is employed to obtain adaptive parameters, using a large-scale external dataset, that can adapt quickly to the novel condition (degradation model) of the given test video during the internal learning task, thereby exploiting external and internal information of a video for super-resolution. The model trained using our approach can quickly adapt to a specific video condition with only a few gradient updates, which reduces the inference time significantly. Extensive experiments on standard datasets demonstrate that our method performs favorably against various state-of-the-art approaches. The project page is available at https://agupt013.github.io/AdaVSR.html

CCS CONCEPTS

• **Computing methodologies** → *Instance-based learning*; *Transfer learning*; *Reconstruction.*

KEYWORDS

Video Super-resolution, Temporal Super-resolution, Meta-Transfer learning, Internal Learning

ACM Reference Format:

Akash Gupta, Padmaja Jonnalagedda, Bir Bhanu, Amit K. Roy-Chowdhury. 2021. Ada-VSR: Adaptive Video Super-Resolution with Meta-Learning. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475320

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3475320

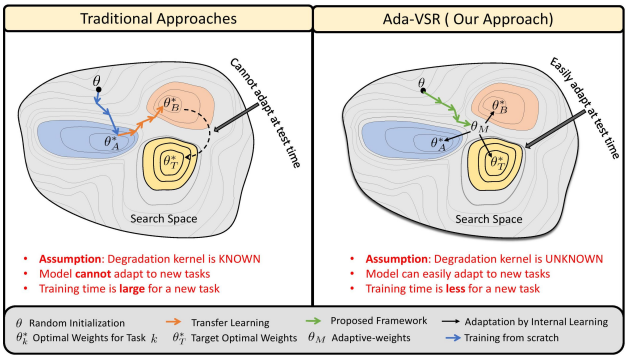


Figure 1: Comparison of traditional approaches against Ada-VSR for the blind VSR task. Traditional supervised approaches train their model assuming the degradation kernel for task A is known (left; blue arrows). Transfer learning can be adopted to find optimal model parameters for Task B with a different degradation kernel (left; orange arrows). However, the model will not be able to generalize for the target task T when degradation is not known. On the other hand, our proposed approach tries to find weights that can easily adapt to the target task with only a few gradient updates via internal learning (right; green arrows. See sec. 3.2)

1 INTRODUCTION

With the increasing popularity of high-performance higher resolution displays such as 4K Ultra HD (UHD), the demand for high-quality visual content is also increasing. However, professional video production and TV screen content are still at Full HD (1080p) resolution [24, 30, 31]. As rendering low-resolution content on higher resolution displays lowers perceptual quality, it calls for improving the resolution of the content to match that of the display. Enhancing the quality of video not only requires increasing the spatial resolution but also the temporal resolution for smooth rendering on high-performance displays. Therefore, it is critical to improve the spatial as well as the temporal resolution of videos to enhance the perceptual quality. Similarly, in microscopy time-lapse imagery, the quality of imaging plays a crucial role in determining the performance of the cell tracking algorithm [9, 13, 58]. Increasing the spatial and temporal resolution of the time-lapse video is desirable to facilitate automatic cell detection and accurate reconstruction of the cell trajectories, respectively. However, capturing the time-lapse video with higher spatio-temporal resolution damages the cell and can lead to disposal of the painstakingly collected experimental data [9, 13, 18].

Most existing approaches have addressed the task of video spatial super-resolution (VSR) [7, 27, 28, 61, 64, 68] and temporal video super-resolution (TSR) [2, 3, 19, 26, 38, 40, 71], separately. A straightforward strategy to perform spatio-temporal video super-resolution

(STVSR) is to cascade the VSR model and the TSR model to generate high-resolution high-frame rate (HR-HFR) video from low-resolution low-frame rate (LR-LFR) video. Nevertheless, this does not yield optimal results as it cannot fully utilize the available spatio-temporal information [67]. Recently, a few works [30, 31, 66, 67], studied the problem of joint spatio-temporal video super-resolution. Zooming Slow-Mo [66] proposed a one-stage STVSR framework using Deformable Convolutional LSTM. The authors in [67] utilize temporal profiles to exploit spatio-temporal information. FISR [31] proposes a multi-scale temporal loss for joint frame-interpolation and super-resolution. However, these approaches require a large dataset of LR-HR pairs with the assumption that the down-sampling kernel to obtain LR frames from HR frames is known and fixed, which does not hold true in a real world setting (Figure 1, left). The problem of blind SR in images, where down-sampling kernel is unknown, is tackled either by estimating the down-sampling kernel [5, 17] or by exploiting the deep internal prior [55, 62] to learn the internal structure of the image. Consequently, these approaches achieve good performance at the expense of heavy computational time as it requires thousands of back-propagation gradient updates for each instance. Another shortcoming of such approaches is that they cannot take advantage of a pre-trained network learned using a large-scale external dataset [57].

Meta-learning has recently garnered much interest to tackle the aforementioned shortcomings. Meta-learning aims to adapt quickly and efficiently to unseen data available at inference time. There are three common approaches to meta-learning: metric-based [56, 60, 63], model-based [42, 47, 51], and optimization-based [14–16]. Model-Agnostic Meta-Learning (MAML) [14] is a gradient-based method and has shown impressive performance by learning the optimal initial state of the model such that it can quickly adapt to a new task with a few gradient steps.

In this paper, we introduce a novel framework **Ada**ptive **V**ideo **S**uper-**R**esolution (**Ada-VSR**) which aims to generate high resolution high-frame rate (HR-HFR) video from a low-resolution low-frame rate (LR-LFR) input. Inspired by meta-transfer learning, we utilize external knowledge as well as internal knowledge from videos for the task of joint spatio-temporal video super-resolution (STVSR). Our approach leverages knowledge from the external dataset and learns adaptive model parameters using meta-learning. As shown in Figure 1 (right), meta-learning is employed to learn initial model parameters that can quickly and efficiently adapt to the test video with unknown degradation. Specifically, we use different down-sampling kernels on a large-scale dataset, as an external learning task, to learn a model that is easy and fast to finetune for novel tasks. The model parameters obtained using external learning allow adaptation to happen more efficiently for fast learning. Next, internal learning is leveraged to finetune the initial model to learn video-instance specific knowledge with limited gradient steps. Since there are only a few gradient steps involved during internal learning for each video, our approach is significantly faster when compared to approaches that completely rely on internal learning and require thousands of gradient updates.

1.1 Approach Overview

An overview of our Adaptive Video Super-Resolution (**Ada-VSR**) training scheme is illustrated in Figure 2. Given a low-resolution low frame-rate input, our objective is to generate a high resolution high frame-rate video when the degradation kernel is unknown. Our **Ada-VSR** approach consists of two networks: the temporal super-resolution module (TSR) denoted by $\mathcal{F}_\theta$ and the spatial super-resolution module (SSR) as $\mathcal{S}_\phi$. We adopt a meta-learning framework to train both the networks jointly using an external dataset containing LR-LFR (obtained using dynamic task generator; refer Fig. 2) and HR-HFR video pairs. The objective of the meta-training is to learn model parameters that can be easily adapted to the test video. This meta-training is performed only once and the adaptive parameters are used to initialize the model in the next step. In a practical setting, we will only have access to the LR-LFR video. Hence, we exploit the internal structure within the test LR-LFR video using internal learning. The objective of internal learning is to finetune the model for the video instance to improve the spatio-temporal super-resolution with only a few gradient steps. Therefore, we first downscale the LR-LFR further to obtain a super low-resolution low-frame rate (SLR-LFR) video. Secondly, we finetune the model to reconstruct the LR-LFR video from the obtained SLR-LFR video to utilize the instance-specific knowledge. As the parameters obtained by external-learning are learnt for fast learning, internal learning allows quick adaptation of the model parameters to blind spatio-temporal super-resolution task. Finally, the trained model is used to infer HF-HFR video from LF-LFR video.

1.2 Contributions

The key contributions of our proposed framework are as follows.

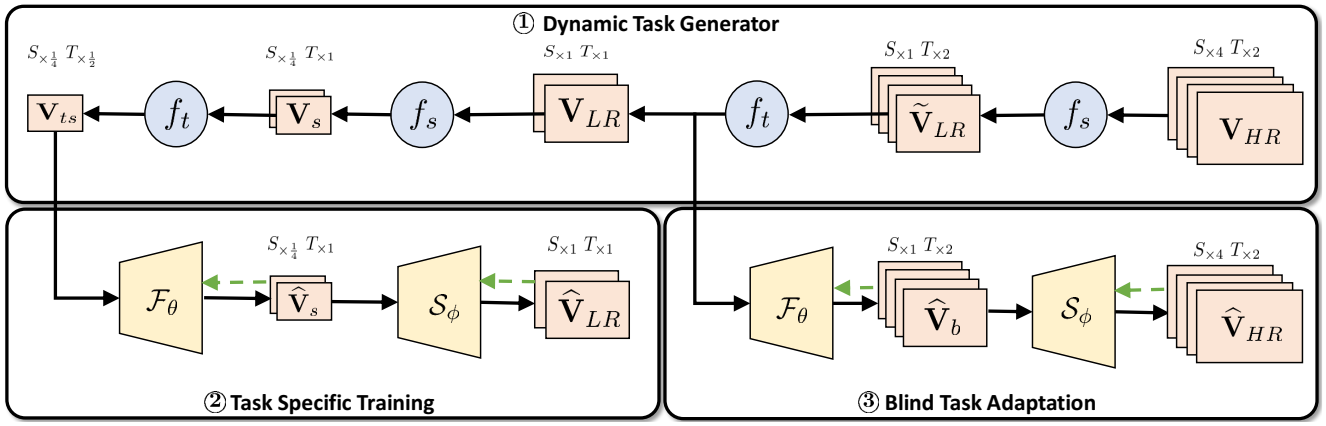
- We propose a novel meta-learning framework **Ada-VSR** for the task of joint spatio-temporal super-resolution by leveraging external and internal learning.
- We employ a combination of various spatial and temporal down-sampling techniques during training to learn a model that can easily adapt to unknown down-sampling/degradation process.
- We significantly reduce the computational time by greatly reducing the gradient steps required during internal learning.

2 RELATED WORK

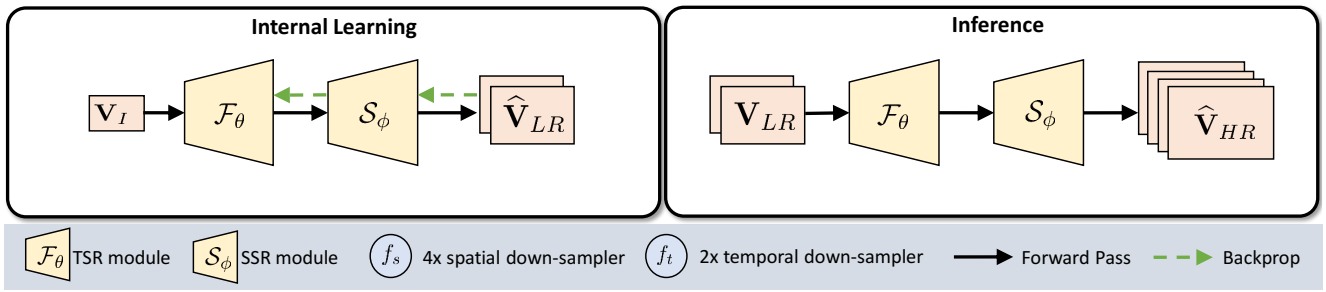
Our work relates to research in spatial and temporal video super-resolution, internal learning and, meta-transfer learning. In this section, we discuss some methods closely related to our work. We provide a characteristic comparison of recent works in Table 1.

Table 1: *Categorization of prior works in video super-resolution.* Different from the state-of-the-art approaches, we employ meta-learning to perform blind spatio-temporal video super-resolution.

Methods	Super-Resolution			Fast Adaptation?
	Spatial	Temporal	Blind	
DyaVSR [33]	✓	✗	✓	✓
Temporal Profiles [67]	✓	✓	✗	✗
Zooming Slow-Mo [66]	✓	✓	✓	✗
Ada-VSR (Ours)	✓	✓	✓	✓



(a) External Learning involves two steps: task-specific training (left) and blind task adaptation (right). Using external dataset $\mathbf{V}_{HR} \in \mathcal{D}_{HR}$ we create a meta-batch with dynamic task generator by spatial down-scaling using f_s and temporal down-scaling using f_t. The meta-batch consists of a train and a test set. First, the train set is used to learn optimal model for task specific training. Then the learnt model is adapted to the blind test task. Following this meta-learning strategy, we obtain parameters which can quickly adapt model during internal learning.



(b) Internal Learning and Inference. Left: We exploit the instance specific information during internal learning. The LR-LFR test video $\mathbf{V}_{LR}$ is used to obtain a spatio-temporal down-sampled video ($\mathbf{V}_I$). For internal learning, the down-sampled video $\mathbf{V}_I$ is used to reconstruct $\widehat{\mathbf{V}}_{LR}$. This helps network to learn the internal statistic of the LR-LFR test video. Right: Finally, HR-HFR video $\widehat{\mathbf{V}}_{HR}$ is generated by passing the test video $\mathbf{V}_{LR}$ through the model trained via internal-learning.

Figure 2: Overview of Adaptive Video Super-Resolution (Ada-VSR) framework. Our framework consists of two modules External Learning and Internal Learning. (a) External learning leverages meta-training protocol and exploits the external dataset to learn parameters that can easily adapt to novel tasks (different degradation in our case). (b) Internal learning helps exploit internal structure of the given video and is used to generate a HF-HFR video ($\widehat{\mathbf{V}}_{HR}$) from LR-LFR video $\mathbf{V}_{LR}$. Since internal learning is initialized by the adaptive parameters obtained through external dataset, our model can quickly adapt to video degraded by unknown kernel with only a few gradient steps.

Image Super-Resolution. Deep learning approaches have shown remarkable performance on the task of image super-resolution [11, 17, 21, 29, 32, 36, 55, 69]. Recently, various Convolutional Neural Network (CNN) based approaches have been proposed for non-blind image SR where the down-sampling kernel (e.g. bicubic), used to obtain low-resolution (LR) image, is known [11, 21, 29, 32, 36]. Despite the impressive performance of these methods, their efficacy deteriorates when the down-sampling kernel is different than the one used to train these models due to the domain gap. To overcome this issue, SRMD [69] incorporates multiple degradation kernels as input to their model along with the LR image. On the other hand, Zero-Shot Super-Resolution (ZSSR) [55] exploits the deep prior [62] to learn an image specific structure to obtain an SR image. Some approaches first try to estimate the degradation kernel and utilize the estimated kernel for image super-resolution. An iterative approach to correct inaccurate degradation kernels is introduced in [17]. Similar to ZSSR, the KernelGAN [5] utilizes the patch-recurrence property of a single image for super-resolution. However, these methods train the network from scratch for all the image instances, making them computationally heavy.

Video Spatial Super-Resolution. Earlier works in video super-resolution focused on developing effective priors on the HR frames to solve this problem [4, 8, 12, 53]. Motivated by the success of deep learning approaches in image super-resolution [17, 55, 69], several deep learning based methods have been proposed for video super-resolution [27, 28, 68]. A CNN based approach is proposed in [28], where the network is trained on both the spatial and temporal dimensions of videos to spatially enhance the frames. An SR draft-ensemble approach for fast video spatial super-resolution is

proposed in [35]. In [7, 61], the authors incorporate optical flow estimation models to explicitly account for the motion between neighboring frames. However, accurate flow is difficult to obtain given occlusion and large motions. A computationally lighter flow estimation module (TOFlow) is proposed in [68] to account for motion information. DUF [27] overcomes this problem by implicit motion compensation using their proposed dynamic upsampling filter network. Pyramid, Cascading and Deformable convolution (PCD) alignment and the Temporal and Spatial Attention (TSA) modules are proposed in EDVR [64] to incorporate implicit motion compensation. However, these approaches assume the degradation kernel for down-sampling is known and/or require a large amount LR-HR pairs to train their models.

Video Temporal Super-Resolution. Video super-resolution can also be performed in the temporal dimension and is often termed as video interpolation. In temporal video super-resolution, the task is to generate a high-frame rate (HFR) video from a low-frame rate video (LFR). Existing approaches [2, 3, 26, 38, 40, 71] use optical flow estimation between input frames for temporal super-resolution. Thus, the quality of estimated optical flow governs the quality of frame interpolation. Deep learning approaches have demonstrated effectiveness in temporal super-resolution tasks. A straightforward application of CNNs for intermediate frame synthesis is presented in [39]. Some methods [45, 46] apply CNNs to estimate space-varying and separable convolutional kernels for frame synthesis using neighbourhood pixels. [1] proposed a non-adversarial approach to generate videos by first learning optimized representation and then interpolating between the optimized latent representation of two frames to synthesize central frame. Joint video deblurring and interpolation to enhance and increase the frame-rate of a video is explored in [19, 54]. These approaches do not perform spatial super-resolution in their work and assume that the low-temporal resolution video is obtained by averaging 9 consecutive frames. Unlike these methods, we address the task of joint spatial and temporal super-resolution.

Meta-Learning. Recently, meta-learning algorithms have achieved impressive performance in various applications like few-shot learning [25, 34, 50, 56, 59], reinforcement learning [14, 20, 44, 52] and image super-resolution [33, 57]. Meta-learning aims to learn a model that can quickly and efficiently adapt to novel unseen tasks. There are three common approaches to meta-learning: metric-based [56, 60, 63], model-based [42, 47, 51], and optimization-based [14–16]. DynaVSR is proposed in [33], which utilizes meta-learning for spatial video super-resolution and has shown superior performance. Different from DynaVSR [33], we leverage meta-learning for the task of joint spatio-temporal video super-resolution.

3 METHODOLOGY

Given a low-resolution low frame-rate video our goal is to generate a high-resolution high frame-rate video in blind video super-resolution setting where the down-scaling kernel is not known at the test time. Let the low-resolution low frame-rate video be denoted by $\mathbf{V}_{LR} = \left[\, \mathsf{L}_1,\, \mathsf{L}_2, \cdots,\, \mathsf{L}_L \right]$, with M frames where $\mathsf{L}_t \in \mathbb{R}^{H \times W \times C}$ and t denotes the time step. We aim to generate a high-resolution high frame-rate video $\mathbf{V}_{HR} = \left[\mathsf{S}_1,\, \mathsf{S}_2, \cdots,\, \mathsf{S}_N\right]$ with N frames, where $\mathsf{S}_t \in \mathbb{R}^{\mathbf{a}H \times \mathbf{a}W \times C}$ and $N = \mathbf{b}M$. Our objective is to increase the spatial resolution of the given input video $\mathbf{V}_{LR}$ by a factor of **a** and the temporal resolution by a factor of **b**.

In this section, we describe the proposed **Ada-VSR** framework in detail. Our framework consists of two modules: the Temporal Super-Resolution (TSR) module $\mathcal{F}_\theta$ and the Spatial Super-Resolution (SSR) module $\mathcal{S}_\phi$. We use the TSR module to interpolate frames and increase the frame rate by a factor of 2. The SSR network uses the output of the temporal super-resolution module and increases the spatial resolution by a scaling factor of 4. The scaling factor values are fixed for all our experiments. The overall training scheme of our approach **Ada-VSR** is shown in Figure 2. Our framework consists of two training paradigms: external learning and internal learning.

3.1 External Learning

The external learning protocol leverages a large-scale external dataset to perform knowledge transfer and domain generalization using pre-training and meta-transfer learning respectively.

Large-Scale Training. In large-scale pre-training, we utilize a high-quality external dataset ($\mathcal{D}_{HR}$) to provide a warm start for meta-transfer learning. Since super-resolution tasks with different down-scaling kernels share similar parameter space, large-scale training helps to estimate the natural prior of high-resolution high-frame rate videos. The large-scale pre-training is also effective to stabilize the training of the meta-learning algorithm MAML [14].

For the SSR module, we apply bi-cubic spatial degradation to HR-HFR video $\mathbf{V}_{HR} \in \mathcal{D}_{HR}$ to obtain low-resolution high-frame rate video LR-HFR $\widetilde{\mathbf{V}}_{LR}$. The videos $\mathbf{V}_{HR}$ and $\widetilde{\mathbf{V}}_{LR}$ form a synthetic dataset $\mathcal{D}_s$. We train the network $\mathcal{S}_\phi$ to learn spatial super-resolution task by minimizing the ℓ_1 reconstruction loss between all the frames of the generated HR-HFR video ($\widehat{\mathbf{V}}_{HR}$) and corresponding ground truth HR-HFR video $\mathbf{V}_{HR}$. The objective function for large-scale training of the SSR module $\mathcal{S}_\phi$ is defined as:

$$\mathcal{L}^{\mathcal{D}_s} = \sum_{(\widetilde{\mathbf{V}}_{LR}, \mathbf{V}_{HR}) \sim \mathcal{D}_s} \left\| \mathcal{S}_\phi(\widetilde{\mathbf{V}}_{LR}) - \mathbf{V}_{HR} \right\|_1 \tag{1}$$

We choose ℓ_1-loss instead of Mean-Squared Error (MSE) ℓ_2 loss as latter has inherent property of generating blurry output as shown in the literature [70].

The TSR module should be able to increase the frame-rate of a low-frame-rate (LFR) video. We can interpolate the frames to increase the frame-rate by factor of 2 and learn a residual by minimizing the reconstruction loss between the generated high-frame-rate $\widehat{\mathbf{V}}_{HR}$ and the ground truth video $\mathbf{V}_{HR}$. However, it may not be able to capture the temporal dynamics efficiently [6, 67]. Recently some works have addressed this by taking temporal profile to leverage the patch recurrence in temporal dimension to train the network efficiently [6, 67]. We adopt the same strategy to train the TSR module $\mathcal{F}_\theta$. We define a temporal profile generator function f_r that takes a video input, performs the bi-cubic interpolation in temporal dimension and returns the temporal profile. To generate a dataset to train the TSR module we select alternate frames of the high-frame-rate (HFR) video $\mathbf{V}_{HR}$ to generate a LFR video $\overline{\mathbf{V}}_{LR}$. Then we apply the temporal profile generator function (f_r) to get the temporal profile $\mathbf{V}'_{LR}$ corresponding to the input $\overline{\mathbf{V}}_{LR}$ such that $\mathbf{V}'_{LR} = f_r(\overline{\mathbf{V}}_{LR})$, where $\mathbf{V}'_{HR}$ is the HFR temporal profile of the LFR input $\overline{\mathbf{V}}_{LR}$. We denote the paired data $(\overline{\mathbf{V}}_{LR}, \mathbf{V}_{HR})$ as $\mathcal{D}_t$. The loss

function to update the TSR module $\mathcal{F}_\theta$ is given below.

$$\mathcal{L}^{\mathcal{D}_t} = \sum_{(\overline{\mathbf{V}}_{LR}, \mathbf{V}_{HR}) \sim \mathcal{D}_t} \left\| \mathcal{F}_\theta(\overline{\mathbf{V}}_{LR}) - \mathbf{V}_{HR} \right\|_1 \tag{2}$$

Dynamic Task Generator The Dynamic Task Generator (DTG; see Fig. 2a) generates tasks for meta-training on-the-fly using diverse degradation settings. In our approach, the task $\mathcal{T}_i$ is the combination of the spatial down-scaling kernel and temporal sub-sampling method. For spatial down-sampling by a factor of 4 we randomly apply the anisotropic Gaussian kernels using the function f_s. Temporal sub-sampling is performed with the function f_t by either selecting alternate frames or by averaging a window of size 3 to obtain a low-frame rate video.

Meta-Transfer Learning. We seek to find a set of transferable initial parameters where a few-gradient steps can adapt the model to the current video and achieve to large performance gain. Motivated by MAML [14] and [57], we employ meta-transfer learning strategy for spatio-temporal video super-resolution (STVSR) to learn adaptive weights. Unlike MAML, we use the external dataset for meta-training and leverage internal learning for meta-test step. Training with external dataset helps the meta-leaner to focus more on the down-scaling kernel-agnostic property, whereas internal learning helps to exploit the instance specific internal statistics.

Lines 10-19 in Algorithm 1 presents the meta-transfer learning optimization protocol. In our approach, we aim to learn a generalized set of TSR parameters θ and SSR parameters ϕ such that the parameters can adapt to the test video quickly and efficiently in a blind super-resolution setting. The meta-learning achieves this using two steps: task-specific training and blind task adaptation.

Task-Specific Training. It is the inner loop of the MAML meta-learning algorithm, the meta-learner tries to learn task-specific optimal parameters in one or more gradient descent steps. The inner loop is represented by Lines 12-16 in Algorithm 1. Given an external dataset $\mathcal{D}_{HR}$, we obtain a meta-task train batch $\mathcal{D}_{tr} = (\mathbf{V}_{LR}, \mathbf{V}_s, \mathbf{V}_{ts})$ for $\mathcal{T}_i \in p(\mathcal{T})$, where $p(\mathcal{T})$ is the task distribution, $\mathbf{V}_{LR}$ is LR-LFR video, $\mathbf{V}_s$ is 4x spatially down-scaled version $\mathbf{V}_{LR}$ and $\mathbf{V}_{ts}$ is 2x temporally down-scaled version of $\mathbf{V}_s$. We train the TSR model ($\mathcal{F}_\theta$) to generate a video $\widehat{\mathbf{V}}_s$ with temporal resolution twice to that of input video ($\mathbf{V}_{ts}$). The SSR model ($\mathcal{F}_\phi$) takes the output of TSR model ($\widehat{\mathbf{V}}_s$) and reconstructs the LR-LFR video $\widehat{\mathbf{V}}_{LR}$. The output of both the models are given by

$$\widehat{\mathbf{V}}_s = \mathcal{F}_\theta(\mathbf{V}_{ts}) \qquad \widehat{\mathbf{V}}_{LR} = \mathcal{S}_\phi(\widehat{\mathbf{V}}_s) \tag{3}$$

We optimize both the networks for n_i iteration to increase the resolution of a video for a task defined by random spatial and temporal down-scaling kernels. The loss function for the task-specific training is computed as follows:

$$\mathcal{L}^{tr}_{\mathcal{T}_i} = \sum_{\mathcal{D}_{tr}} \sum_{n_i} \left(\mathcal{L}\left(\widehat{\mathbf{V}}_s, \mathbf{V}_s\right) + \mathcal{L}\left(\widehat{\mathbf{V}}_{LR}, \mathbf{V}_{LR}\right) \right) \tag{4}$$

where $\mathcal{L}$ is reconstruction loss, n_i is number of inner loop iterations for task-specific training. For one gradient update, new adapted parameters θ_i and ϕ_i are then obtained as

$$\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi) \tag{5}$$

$$\phi i = \phi - \alpha \nabla_\phi \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi) \tag{6}$$

where α is the task-level learning rate.

Algorithm 1: Ada-VSR External Training

Input : High-resolution high-frame rate dataset $\mathcal{D}_{HR}$ and task distribution $p(\mathcal{T})$
Input : α, β: task-specific and adaptation learning rate
Output : **Ada-VSR** model parameters θ and ϕ

/* Large-Scale training */
1 Randomly initialize θ, ϕ
2 Generate $\mathcal{D}_s$ using bi-cubic down-sampling kernel on $\mathcal{D}_{HR}$
3 **while** *not done* **do**
  /* Train SSR module */
4 Sample LR-HR batch from $\mathcal{D}_s$
5 Compute $\mathcal{L}^{\mathcal{D}_s}$ by Eq. (1)
6 Update ϕ with respect to $\mathcal{L}^{\mathcal{D}_s}$
  /* Train TSR module */
7 Sample LFR-HFR batch from $\mathcal{D}_t$
8 Compute $\mathcal{L}^{\mathcal{D}_t}$ by Eq. (2)
9 Update θ with respect to $\mathcal{L}^{\mathcal{D}_t}$
/* Meta-Transfer Learning */
10 **while** *not done* **do**
11 Sample task batch $\mathcal{D}_{tr}$, $\mathcal{D}_{te}$ for the task $\mathcal{T}_i, \mathcal{T}_j \sim p(\mathcal{T})$
  /* Task-Specific Training (inner loop) */
12 **for** *all* $\mathcal{T}_i$ **do**
13 Compute meta-training loss ($\mathcal{D}_{tr}$): $\mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi)$
14 Adapt parameters with gradient descent:
15 $\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi), \quad \phi_i = \phi - \beta \nabla_\phi \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi)$
  /* Blind Task Adaptation (outer loop) */
16 Update θ and ϕ with respect to average test loss ($\mathcal{D}_{te}$):
17 $\theta \leftarrow \theta - \alpha \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_i}(\theta_i, \phi_i)$
18 $\phi \leftarrow \phi - \beta \nabla_\phi \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_i}(\theta_i, \phi_i)$

Algorithm 2: Ada-VSR Internal Learning

Input : LR-LFR test video $\mathbf{V}_{LR}$, meta-transfer trained model parameter θ, ϕ, number of gradient updates n and learning rate γ
Output : High-resolution high-frame rate video $\widehat{\mathbf{V}}_{HR}$
1 Generate down-sampled video $\mathbf{V}_I$ by down-sampling $\mathbf{V}_{LR}$ with corresponding blur kernel.
/* Internal Learning */
2 **for** *n steps* **do**
3 Evaluate loss $\mathcal{L}_{int}(\theta, \phi)$ using (10)
4 Update $\theta \leftarrow \theta - \gamma \nabla_\theta \mathcal{L}_{int}(\theta, \phi)$
5 Update $\phi \leftarrow \phi - \gamma \nabla_\phi \mathcal{L}_{int}(\theta, \phi)$
/* Inference to generate HR-HFR */
6 **return** $\widehat{\mathbf{V}}_{HR} = \mathcal{S}_\phi\left(\mathcal{F}_\theta(\mathbf{V}_{LR})\right)$

Blind Task Adaptation. The blind task-adaptation is the outer loop of meta-learning which adapts the model parameters to the novel task. Here the meta-test batch $\mathcal{D}_{te}$ is sampled from $\mathcal{D}_{HR}$ such that $\mathcal{D}_{te} = (\mathbf{V}_{HR}, \widetilde{\mathbf{V}}_{LR}, \mathbf{V}_{LR})$ for $\mathcal{T}_j \in p(\mathcal{T})$ where $\mathcal{T}_i \neq \mathcal{T}_j$, $\mathbf{V}_{HR}$ is HR-LFR video, $\widetilde{\mathbf{V}}_{LR}$ is 4x spatially down-scaled version of $\mathbf{V}_{HR}$

and $\mathbf{V}_{LR}$ is 2x temporally down-scaled version of $\widetilde{\mathbf{V}}_{LR}$. In order to adapt the models to new task $\mathcal{T}_j$, the model parameters θ and ϕ are optimized to achieve minimal test error on $\mathcal{D}_{te}$ with respect to θ_i and ϕ_i. The meta-objective for blind task-adaptation is

$$\begin{aligned} &\arg\min_{\theta,\phi} \sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta_i, \phi_i) \\ &= \arg\min_{\theta,\phi} \sum_{\mathcal{T}_j} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta - \alpha\nabla_\theta \mathcal{L}^{tr}_{\mathcal{T}_i},\ \phi - \alpha\nabla_\phi \mathcal{L}^{tr}_{\mathcal{T}_i}) \end{aligned} \tag{7}$$

Blind task adaptation using equation (8) learns the knowledge across tasks $\mathcal{T}_i$ and $\mathcal{T}_j$. The parameter update rule for for the above optimization can be expressed as:

$$\theta \leftarrow \theta - \beta\nabla_\theta \sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta_i, \phi_i) \tag{8}$$

$$\phi \leftarrow \phi - \beta\nabla_\phi \sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta_i, \phi_i) \tag{9}$$

where β is the learning rate for blind task adaptation step.

3.2 Internal Learning and Inference

Algorithm 2 presents the internal learning and inference steps of our proposed approach. Given a LR-LFR video, we spatially down-sample it with corresponding down-sampling kernel by adopting the kernel estimation algorithms in [41, 48] for blind scenario and select alternate frames from the LR-LFR video to generate $\mathbf{V}_I$ and perform a few gradient updates with respect to the model parameter using a single pair of $\mathbf{V}_I$ as input and a given LR-LFR video $\mathbf{V}_{LR}$ as ground truth (Algorithm 2 Line 2-5). The aim here is to learn the internal statistics of the given video which can be utilized while generating HR-HFR video during inference. The objective function for internal learning is given:

$$\mathcal{L}_{int} = \left\| S_\phi\big(\mathcal{F}_\theta(\mathbf{V}_I)\big) - \mathbf{V}_{LR} \right\|_1 \tag{10}$$

Then, we use the model trained with internal learning for inference. We feed the given LR-LFR input video $\mathbf{V}_{LR}$ to the model to generate a HR-HFR video $\widehat{\mathbf{V}}_{HR}$ as shown in Algorithm 2 Line 6.

4 EXPERIMENTS

In this section, we first introduce the benchmark datasets and evaluation metrics. The qualitative and quantitative experiments are shown to demonstrate the effectiveness of our proposed approach in generating high-resolution high frame-rate videos.

4.1 Datasets and Metrics

We evaluate the performance of our approach using publicly available Vimeo-90K [68] and Vid4 [37] datasets which have been used in many prior spatial video super-resolution and temporal super-resolution works.

Vimeo-90K Dataset. The Vimeo-90K [68] dataset contains 91,707 short video clips, each containing 7 frames. The spatial resolution of each frame is (448 × 256). We use Vimeo-90K only for pre-training and meta-training, using the training split of 64,612 clips and use to the test set to compare against state-of-the-art approaches.

Vid4 Dataset. The Vid4 dataset [37],contains four video sequences: city, walk, calendar, and foliage. All the videos in Vid4 dataset contain at least 30 frames each and are of spatial resolution 720×480. We evaluate a model trained on Vimeo-90K dataset on the Vid4 dataset and report the performance.

Metrics. For quantitative evaluation, we compare three metrics that evaluate different aspects of output image quality: Peak Signal-to-Noise Ratio (PSNR) [23], Structural Similarity Index Measure (SSIM) [65] and Naturalness Image Quality Evaluator (NIQE) [43].

4.2 Implementation Details

Our framework is implemented in PyTorch [49]. All the experiments are trained with a batch size of 32. We employ ADAM optimizer as the meta-optimizer in the meta-transfer learning step. The task-specific learning rate α is set to 0.01 and the adaptation learning rate β is set to 0.0001 for all our training experiments. The number of iterations in the task-specific training n_i is set to 10. We extracted training patches with a size of 64×64 for large-scale training. We utilize the Vimeo-90K dataset train split as the external dataset for large-scale training and meta-transfer learning. For internal learning the learning rate γ is set to 0.0001.

4.3 Qualitative Results.

Figure 3 compares a high-resolution high frame-rate videos generated using the proposed **Ada-VSR** approach with other state-of-the-art methods given a low-resolution low frame-rate video (left column). The low-resolution low-frame-rate input video is obtained by applying a non-bicubic degradation to the alternate frame of a high-resolution high-frame rate video. The skipped frames are faded in the Figure 3. It can be seen that the temporal profile based approach [67] does not perform well. It is due the fact that the model is trained with the assumption that there is a bi-cubic relationship between the LR-LFR and HR-HFR videos. This assumption is violated when we use an input video which was obtained by a non-bicubic down-sampling. Zooming Slow-Mo [66] produces slightly better video compared to the temporal profile approach as it is able to exploit the internal structure within the video. However, it is not able to exploit the external knowledge and the output is still blurry. Our proposed approach, **Ada-VSR**, produces higher-quality and more visually appealing output as compared to both of these approaches. The performance of our approach can be attributed to the adaptive parameters learned on external dataset with meta-learning. These parameters provide good initial parameters for internal training to learn instance specific characteristics.

4.4 Quantitative Results

Our proposed method performs joint spatio-temporal video super-resolution. We compare **Ada-VSR** against representative STVSR approaches. We first compare our work with the two-stage solutions by cascading a temporal super-resolution module (TSR) and a spatial super-resolution module (SSR). For temporal super-resolution module (TSR) we select SepConv [46] and DAIN [2] model, while SAN [10], IMDN [22], DynaVSR [33], and EDVR [64] are selected for spatial video super-resolution module (SSR). We also compare our work with recently proposed one-stage STVSR Zooming Slow-Mo [66] and Temporal profile based approach [67] where spatio-temporal super-resolution is performed jointly.

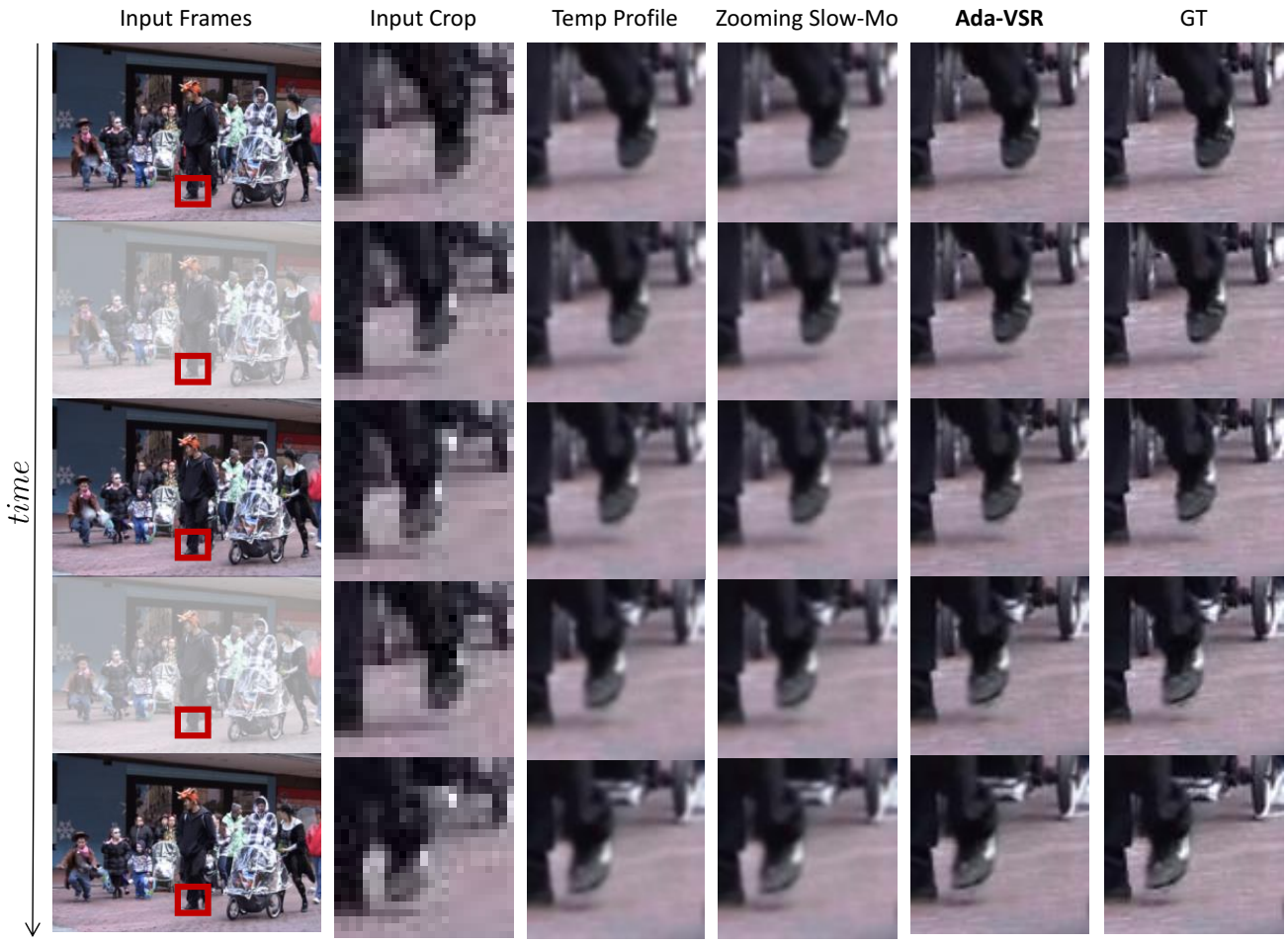


Figure 3: Comparison of qualitative results with the state-of-the-art methods. First column consists of the low-resolution low-frame-rate video input with non-bi-cubic degradation and the missing frames are faded. Second column and last column are the input and ground-truth crop of the input frame region marked in red. As opposed to temporal profile approach, Zooming Slow-Mo and Ada-VSR perform better as they can exploit internal structure of the input. Ada-VSR produces visually appealing results than Zooming Slow-Mo as the weights can easily adapt to novel tasks. Zoom-in for better visualization.

Table 2 presents the quantitative comparison on the test-set of Vimeo-90K [68] dataset and Vid4 [37] dataset. It is evident that our approach outperforms all the two-stage approaches by a significant margin against all the three metrics. When compared with the state-of-the-art one-stage approaches, temporal profile based [67] and Zooming Slow-Mo [66], our proposed **Ada-VSR** achieves superior performance on both the dataset except on Vimeo-90K Slow dataset in terms of PSNR where our performance is only 0.04db less than the temporal profile based approach.

In Table 3, we compare the average inference time of different state-of-the-art approaches. As our method learns adaptive weights that can easily adapt to novel tasks, only a few gradient updates during internal learning step are required to achieve visually compelling results. It can be observed from Table 3 that we outperform the Zooming Slow-Mo [66] by a margin of 0.66dB and the temporal profile approach by a significant margin of +1.18dB. As mentioned earlier, the temporal profile approach assumes that the degradation is bi-cubic hence it cannot generalize well on videos degraded using different blur kernels. It should also be noted that our approach is at about twice as fast as the temporal profile approach and at least 3 times faster when compared with Zooming Slow-Mo. Thus, significantly reducing the inference time during for new test videos with unknown degradation. Values for Zooming Slow-Mo [66] and temporal-profile [67] are reported from [67].

4.5 Ablation Study

We investigate the contribution of large-scale training of the TSR module ($\mathcal{F}_\theta$) and the SSR module ($\mathcal{S}_\phi$). First, for the task-specific training in meta-learning, only the SSR module is initialized using the pre-trained weights obtained in the large-scale training step and the TSR module is initialized randomly. In second case, only the TSR module is initialized with the weights obtained from the

Table 2: Quantitative results comparison on Vimeo-90K [68] and Vid4 [37] datasets. Our proposed approach is very competitive against various state-of-the-art approaches. Best scores are shown in bold and the second best are underlined. The state-of-the-art results are reported from [67].

Method		*Vimeo-90K Slow*			*Vimeo-90K Medium*			*Vimeo-90K Fast*			*Vid4*		
TSR	**SSR**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**
SepConv [46]	IMDN [22]	31.75	0.88	7.68	33.13	0.90	7.78	34.31	0.92	8.55	24.87	0.72	6.34
SepConv [46]	SAN [10]	32.12	0.90	7.10	33.59	0.91	7.46	34.97	0.92	8.48	24.93	0.72	5.89
SepConv [46]	EDVR [64]	32.97	0.91	7.00	34.25	0.92	7.40	35.51	0.92	8.48	25.93	0.78	5.70
DAIN [2]	IMDN [22]	31.84	0.89	7.13	33.39	0.91	7.58	34.74	0.92	8.43	24.93	0.72	6.18
DAIN [2]	SAN [10]	32.26	0.90	7.05	33.82	0.92	7.45	35.27	0.92	8.48	25.14	0.73	5.78
DAIN [2]	EDVR [64]	33.21	0.91	7.06	34.73	0.93	7.39	35.71	0.93	8.47	26.12	0.79	5.62
SSR	**TSR**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**
IMDN [22]	SepConv [46]	32.01	0.89	7.67	33.22	0.90	7.65	34.50	0.92	8.54	24.88	0.72	6.33
IMDN [22]	DAIN [2]	32.27	0.89	6.99	33.73	0.92	7.17	35.15	0.92	8.41	24.99	0.72	6.2
SAN [10]	SepConv [46]	32.32	0.90	6.99	33.73	0.92	7.32	35.33	0.92	8.42	25.01	0.73	5.87
SAN [10]	DAIN [2]	32.56	0.91	6.90	34.12	0.93	7.43	35.47	0.92	8.39	25.26	0.75	6.16
DynaVSR [33]	DAIN [2]	-	-	-	-	-	-	-	-	-	26.54	0.81	5.65
End-to-end Framework		**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**
Zooming Slow-Mo [66]		33.29	0.91	6.94	35.24	0.93	7.35	36.43	0.93	8.41	26.30	0.80	5.62
Temporal Profile [67]		**33.40**	**0.92**	6.17	35.55	0.94	6.37	36.29	0.93	7.13	26.50	0.82	5.48
Ada-VSR (Ours)		33.36	**0.92**	**6.12**	**35.91**	**0.95**	**6.33**	**36.52**	**0.95**	**6.99**	**26.98**	**0.84**	**5.40**

Table 3: Average inference time (sec per frames) comparison of Ada-VSR with recent approaches for blind spatio-temporal video super-resolution.

Method	*Vid4 [37]*	
	PSNR↑	**Avg. time↓**
Zooming Slow-Mo [66]	26.30	0.1995
DynaVSR [33] + DAIN [2]	26.54	0.8940
Temporal profile [67]	25.78	0.1328
Ada-VSR (Ours)	**26.96**	**0.0680**

Table 4: Impact of large-scale training of TSR and SSR modules in Ada-VSR on the target performance.

External-Training		*Vid4 [37]*		
Spatial	**Temporal**	**PSNR ↑**	**SSIM↑**	**NIQE↓**
✓	✗	25.98	0.80	5.77
✗	✓	26.27	0.81	5.59
✓	✓	**26.98**	**0.84**	**5.40**

large-scale training and the SSR module is randomly initialized. The quantitative results of impact of large-scale training using external data on Vid4 dataset are shown in Table 4. We can observe that the performance of **Ada-VSR** drops if large-scale pre-training is performed on only one of the SSR or TSR module. This is expected since the meta-learning algorithm MAML [14] has shown to be unstable when training without a warm model initialization. It is interesting to note that the model with large-scale training of only TSR module outperforms the one initialized with only the SSR module. We believe this could be due to the rich information available in temporal profiles used for large-scale training of the TSR module which provide stable initial parameters for task-specific training and blind-task adaptation during meta-training.

5 CONCLUSIONS

We present an Adaptive Video Super Resolution framework (**Ada-VSR**) for generating high resolution high frame-rate videos from low resolution low frame-rate input videos. We leverage external as well as internal learning to achieve spatio-temporal super-resolution. Specifically, external learning employs meta-learning to learn adaptive network parameters that can easily adapt unknown degradation, while internal learning, on the other hand, helps to capture the underlying statistics of down-sampling and degradation specific to the input video by exploiting the internal structure, thereby making our approach more suited for practical, real-world data enhancement tasks. The proposed approach is able to achieve superior enhancement while adapting to unknown degradation models as shown in our experiments. Experiments on standard datasets show not only the quantitative and qualitative efficacy of our proposed model in joint spatio-temporal video super-resolution, but also the improvement in computational time over various state-of-the-art methods.

ACKNOWLEDGMENTS

The work was partially supported by NSF grants 1664172, 1724341 and 1911197.

REFERENCES

[1] Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M Salman Asif, and Amit K Roy-Chowdhury. 2020. Non-Adversarial Video Synthesis with Learned Priors. *arXiv preprint arXiv:2003.09565* (2020).

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. Depth-aware video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3703–3712.

[3] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[4] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. 1996. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*. Springer, 571–582.

[5] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. 2019. Blind super-resolution kernel estimation using an internal-gan. *arXiv preprint arXiv:1909.06581* (2019).

[6] Gordon J Berman. 2018. Measuring behavior across scales. *BMC biology* 16, 1 (2018), 1–11.

[7] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4778–4787.

[8] Stephen C Cain, Majeed M Hayat, and Ernest E Armstrong. 2001. Projection-based image registration in the presence of fixed-pattern noise. *IEEE transactions on image processing* 10, 12 (2001), 1860–1872.

[9] MC Comes, P Casti, A Mencattini, D Di Giuseppe, F Mermet-Meillon, A De Ninno, MC Parrini, L Businaro, C Di Natale, and E Martinelli. 2019. The influence of spatial and temporal resolutions on the analysis of cell-cell interaction: a systematic study for time-lapse microscopy applications. *Scientific reports* 9, 1 (2019), 1–11.

[10] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11065–11074.

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* 38, 2 (2015), 295–307.

[12] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. 2004. Fast and robust multiframe super resolution. *IEEE transactions on image processing* 13, 10 (2004), 1327–1344.

[13] Gianmarco Ferri, Luca Digiacomo, Francesca D'Autilia, William Durso, Giulio Caracciolo, and Francesco Cardarelli. 2018. Time-lapse confocal imaging datasets to assess structural and dynamic properties of subcellular nanostructures. *Scientific data* 5, 1 (2018), 1–8.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.

[15] Chelsea Finn and Sergey Levine. 2017. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *arXiv preprint arXiv:1710.11622* (2017).

[16] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. 2018. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930* (2018).

[17] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. 2019. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1604–1613.

[18] Akash Gupta, Abhishek Aich, Kevin Rodriguez, G Venugopala Reddy, and Amit K Roy-Chowdhury. 2020. Deep Quantized Representation For Enhanced Reconstruction. In *2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)*. IEEE, 1–4.

[19] Akash Gupta, Abhishek Aich, and Amit K Roy-Chowdhury. 2020. ALANET: Adaptive Latent Attention Network for Joint Video Deblurring and Interpolation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 256–264.

[20] Abhishek Gupta, Benjamin Eysenbach, Chelsea Finn, and Sergey Levine. 2018. Unsupervised meta-learning for reinforcement learning. *arXiv preprint arXiv:1806.04640* (2018).

[21] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. 2018. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1664–1673.

[22] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. 2019. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2024–2032.

[23] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* 44, 13 (2008), 800–801.

[24] Posted in Video Animation. 2018. https://www.introbrand.com/blog/all-you-need-to-know-about-video-resolutions-and-frame-rates/ Accessed: 2021-04-04.

[25] Muhammad Abdullah Jamal and Guo-Jun Qi. 2019. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11719–11727.

[26] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9000–9008.

[27] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3224–3232.

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

[29] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1646–1654.

[30] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2019. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3116–3125.

[31] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2020. FISR: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11278–11286.

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[33] Suyoung Lee, Myungsub Choi, and Kyoung Mu Lee. 2021. DynaVSR: Dynamic Adaptive Blind Video Super-Resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2093–2102.

[34] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835* (2017).

[35] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. 2015. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 531–539.

[36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 136–144.

[37] Ce Liu and Deqing Sun. 2011. A bayesian approach to adaptive video super resolution. In *CVPR 2011*. IEEE, 209–216.

[38] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. 2017. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*. 4463–4471.

[39] Gucan Long, Laurent Kneip, Jose M Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. 2016. Learning image matching by simply watching video. In *European Conference on Computer Vision*. Springer, 434–450.

[40] Dhruv Mahajan, Fu-Chung Huang, Wojciech Matusik, Ravi Ramamoorthi, and Peter Belhumeur. 2009. Moving gradients: a path-based method for plausible image interpolation. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–11.

[41] Tomer Michaeli and Michal Irani. 2013. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 945–952.

[42] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141* (2017).

[43] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.

[44] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. 2018. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347* (2018).

[45] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 670–679.

[46] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 261–270.

[47] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123* (2018).

[48] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. 2016. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1628–1636.

[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in PyTorch. In *NIPS AutoDiff Workshop*.

[50] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676* (2018).

[51] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*. PMLR, 1842–1850.

[52] Nicolas Schweighofer and Kenji Doya. 2003. Meta-learning in reinforcement learning. *Neural Networks* 16, 1 (2003), 5–9.

[53] Oded Shahar, Alon Faktor, and Michal Irani. 2011. *Space-time super-resolution from a single video*. IEEE.

[54] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. 2020. Blurry Video Frame Interpolation. arXiv:2002.12259 [cs.CV]

[55] Assaf Shocher, Nadav Cohen, and Michal Irani. 2018. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3118–3126.

[56] Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175* (2017).

[57] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. 2020. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3516–3525.

[58] Yu-Ting Su, Yao Lu, Jing Liu, Mei Chen, and An-An Liu. 2021. Spatio-Temporal Mitosis Detection in Time-Lapse Phase-Contrast Microscopy Image Sequences: A Benchmark. *IEEE Transactions on Medical Imaging* 40, 5 (2021), 1319–1328.

[59] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 403–412.

[60] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1199–1208.

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

[62] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9446–9454.

[63] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080* (2016).

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

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

[66] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. 2020. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3370–3379.

[67] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. 2020. Space-Time Video Super-Resolution Using Temporal Profiles. In *Proceedings of the 28th ACM International Conference on Multimedia*. 664–672.

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

[69] Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2018. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3262–3271.

[70] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2015. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861* (2015).

[71] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2004. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)* 23, 3 (2004), 600–608.