IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A novel Zero-Shot Real World Spatio-Temporal Super-Resolution (ZS-RW-STSR) model for Video Super-Resolution

ANKIT SHUKLA[1], AVINASH UPADHYAY[1], MANOJ SHARMA[1], ANIL SAINI [2], NUZHAT FATEMA [3], HASMAT MALIK[4], ASYRAF AFTHANORHAN[3], MOHAMMAD ASEF HOSSAINI[5]

[1]Bennett University, Greater Noida, India
[2]CSIR-CEERI, Pilani, India
[3]Faculty of Bussiness and Management, Universiti Sultan Abidin (UniSA), Terengganu, Malaysia
[4]Deparment of Electrical Power Engineering, Faulty of Electrical Engineering, Universiti Teknologi Malaysia, Johor Bahru, 81310
[5]Deparment of Physics, Badghis University, Afganistan

Corresponding Author: MOHAMMAD ASEF HOSSAINI (e-mail: asef.hossaini_edu@basu.edu.af)

ABSTRACT Super-resolution (SR) of the degraded and real low-resolution (LR) video remains a challenging problem despite the development of deep learning-based SR models. Most existing state-of-the-art networks focus on getting high-resolution (HR) videos from the corresponding down-sampled LR video but fail in scenarios with noisy or degraded low-resolution video. In this article, a novel real-world "zero-shot" video spatio-temporal SR model, i.e., 3D-Deep Convolutional Auto-Encoder (3D-CAE) guided attention-based deep spatio-temporal back-projection network has been proposed. 3D-CAE is utilized for extracting noise-free features from real low-resolution video and used in the attention-based deep spatio-temporal back-projection network for clean, high-resolution video reconstruction. In the proposed framework, the denoising loss of low-resolution video with high-resolution video reconstruction loss is jointly used in an end-to-end manner with a zero-shot setting. Further, Meta-learning is used to initialize the weights of the proposed model to take advantage of learning on the external dataset with internal learning in a zero-shot environment. To maintain the temporal coherency, we have used the Motion Compensation Transformer (MCT) for motion estimation and the Sub-Pixel Motion Compensation (SPMC) layer for motion compensation. We have evaluated the performance of our proposed model on REDS and Vid4 Dataset. The PSNR value of our model is 25.13 dB for the RealVSR dataset, which is 0.72 dB more than the next-best performing model, EAVSR+. For MVSR4x, our model provides 24.61 db PSNR, 0.67 dB more than the EAVSR+ model. Experimental results demonstrate the effectiveness of the proposed framework on degraded and noisy real low-resolution video compared to the existing methods. Furthermore, an ablation study has been conducted to highlight the contribution of 3D-CAE and attention layer to the overall network performance.

INDEX TERMS Zero-shot, Super-Resolution, and Convolutional Auto-encoder.

I. INTRODUCTION

Videos are the most common and comprehensive source of information in today's day-to-day life. With the advent of high-tech imaging technologies, videos can be captured in HD and UHD quality, thereby enhancing one's perceptual experience. However, with such technologies, there are certain situations (remote sensing [1]–[3], UAV surveillance, etc.) where capturing HD videos is complex or involves more cost. The high-resolution cameras in these cases are expensive, and transmitting such media requires enormous bandwidth.

Video super-resolution (VSR) is a computational technique that intends to address these challenges by generating a high-resolution (HR) sequence of video frames from corresponding low-resolution (LR) video frames. The VSR [4] has numerous applications in the fields of remote sensing, UAV surveillance, Panorama video super-resolution, secu-

rity, High Definition(HD) and Ultra HD (UHD) Televisions, etc.

Though various techniques have evolved for single image super-resolution (SISR) [5]–[8], the VSR is still a demanding and ill-posed problem. Contrary to SISR, where a single image is super-resolved, VSR intends to encapsulate the inter-frame alignments while performing frame-to-frame super-resolution.

The VSR method has two domains: spatial and temporal SR. The spatial SR intends to increase the size of the frames while preserving-cum-adding additional information. In contrast, temporal SR [9] intends to reduce the quantization loss of information between two frames. Temporal SR is the retrieval of those dynamic events that occur faster than the provided frame rate by predicting mid-frame information. These pieces of inter-frame information are critical in VSR to maintain consistency in the motion of the video. Space-time video SR is more challenging as spatial and temporal SR are both ill-posed problems [10]. This problem is more interesting and valuable in many computer vision and biomedical tasks for pre-processing of videos. The problem is further complicated when there is degradation in the video frames.

Recently, Deep learning based VSR models have provided state-of-the-art (SOTA) performance but have limitations such as high computational complexity, dataset-specific performance [11], and adaptation to synthetic LR degradations (such as bicubic, etc.) exclusively [12]. These frameworks were not generalized and were trained over synthetically generated LR sequences. Hence, they performed well over the dataset with synthetic LR video sequences. Still, their performance deteriorated significantly when the video frame sequences from other datasets having real degradation were evaluated [13]. The real noises are heterogeneous, with different degradations having no verifiable mathematical models. The generation of HR video frame sequences from the dataset having such noises is extremely difficult. To address the issues of real-world noise scenarios in SISR-based models, the ZSSR [14], a self-supervised method, used a zero-shot setting to learn the internal information (non-local structures) of the image on a simple CNN model. The model outperformed different blur kernels compared to other SISR SOTA models with minimal computational complexity. However, it could not exploit the patterns of large external datasets, resulting in a non-generalized and less adaptive model. It took thousands of iterations to learn the information in the sample before it could produce good results. This problem was overcome using meta-learning in MZSR [15]. In MZSR [15], they first trained the model on an external large dataset, a transfer learning step, and then they meta-trained the model over different blur kernels to incorporate kernel-agnostic characteristics in the model. With these weights, they used zero-shot training over the LR image to produce the SR Image. The model was able to exploit the information of the large external dataset and internal non-local structures of the images to produce exceptionally good and generalized results with faster learning. The model was able to adapt to new samples with different blur kernels in a few gradient descent steps as compared to ZSSR [14].

Inspired by the success of these results on SISR problems, in this article, a novel zero-shot and meta-learning-based real-world video space-time SR method (3D-deep **C**onvolutional **A**uto-encoder guided attention-based deep **S**patio- **T**emporal back-projection **Net**work (CASTNet)) is introduced for super-resolution of real LR videos. The 3D Convluational auto-encoder (3D-CAE) learns the noise-free features from noisy LR videos, and the Attention-based Deep Spatio-Temporal back-projection Network generates the HR video from these noise-free features. The last three layers of the 3D-CAE are concatenated, and these concatenated features are fed to the deep spatio-temporal network. The spatio-temporal SR network upscaled the features in the spatial and temporal domain and reconstructed the HR video using these up-scaled features. The denoising and SR loss gradients were used to update the denoising model weights and SR weights were updated by the gradient of SR loss to get the improved resultant HR video. Both models were jointly trained in the end-to-end fashion. As given in Figure 1, the model was initially trained on a large dataset, where various blur kernels were used to create different tasks, and these tasks were learned in a meta-learning fashion to update the weights of the model and reduce its kernel dependency. After the learning on the external dataset, for the Meta-test phase, the model was updated using a zero-shot setting for a sample of video frame sequence to generate video-specific SR. Detailed experimental results demonstrate the effectiveness of the proposed meta-learning and zero-shot-based video SR framework on degraded and noisy real low-resolution video compared to the existing methods. Furthermore, an ablation study has been conducted to highlight the contribution of each component of the proposed network. The enhancement module is referred to as ADST-BPN in this article.

II. RELATED WORK

A. IMAGE SUPER-RESOLUTION

Recent advancements in image super-resolution (SR) have been driven by deep learning techniques like [5], [7], [8], [16]–[21]. While effective, these methods rely on knowing the exact degradation kernel, posing limitations in real-world applications. To overcome this, blind image super-resolution has emerged, focusing on self-supervised estimation of unknown degradation kernels. This approach categorizes methods into non-blind SR and blind SR, offering solutions for scenarios where precise kernel information is unavailable.

Non-blind SR methods use the known degradation kernel to generate high-quality, high-resolution (HR) images. Examples include SRM [22], which uses the low-resolution (LR) image and its corresponding degradation kernel as inputs, and ZSSR [14], which trains an image-specific network on the pseudo-LR-HR image-pairs obtained using the same

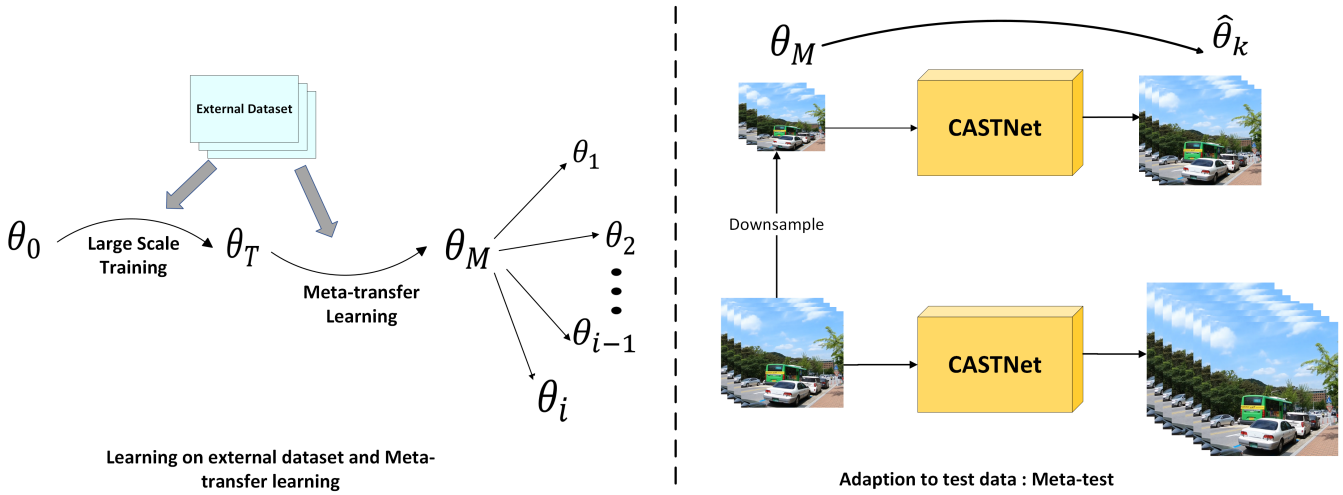

FIGURE 1: The proposed methodology. In the first stage, the external dataset is used for large-scale training. From the initial point, the θ_0 model is trained to obtain θ_T using large-scale dataset REDS. Then during meta-transfer learning θ_M is obtained for various blur kernels and real-world degradations. During the meta-test phase, a test input is downsampled and the model is trained using a self-supervised internal training mechanism

kernel which used to generate LR image from the test image itself.

In contrast to supervised methods, blind super-resolution (SR) techniques aim to infer unknown kernels through self-supervision and subsequently apply these estimated kernels to non-blind SR models. Various strategies have been developed for kernel estimation, leveraging self-similarity or employing iterative self-correction mechanisms. The pioneering work by Michaeli and Irani [23] introduced a method to estimate downscaling kernels by exploiting the patch-recurrence characteristic within a single image. Building upon this, KernelGAN [24] enhanced kernel estimation by incorporating Internal-GAN. Additionally, IKC [25] proposed an iterative correction approach, demonstrating its efficacy in producing high-fidelity SR images.

Also, the feedback mechanism is exploited by Li, Zhen, et al. [26] to refine the output of the network. Another mechanism is proposed in FENet by Behjati et al. [27] using a frequency-based enhancement network. Luo et al. [28] proposed a novel adversarial neural degradation (AND) model for blind image SR to generate a wide range of complex degradation effects that are highly non-linear.

In the context of blind image super-resolution (SR), self-supervised methodologies have been put forth [29], [30]. Dong et al. [29] introduced a self-supervised technique that estimates the blur kernel and intermediary high-resolution (HR) image from a single low-resolution (LR) input image. This approach employs a variational model, grounded in the image formation of SR, to enhance the quality of the intermediary HR images. A separate self-supervised method [30] has integrated contrastive learning into blind remote sensing image SR, directing the reconstruction process by promoting positive representations and penalizing negatives.

Recently, diffusion-based techniques have attracted significant attention in the field of image SR. One such method, SinSR [31], accomplishes single-step SR generation through the derivation of a deterministic sampling process from the most recent state-of-the-art (SOTA) method, thereby expediting diffusion-based SR. Another diffusion-based SR method, EDiffSR [32], utilizes a diffusion probabilistic model, incorporating an Efficient Activation Network (EANet) for enhanced noise prediction performance and a Conditional Prior Enhancement Module (CPEM) for precise super-resolution. Guo et al. [33] have proposed a face video SR method that addresses video compression artifacts by capitalizing on the correlation among video, audio, and the emotional state of the face.

B. VIDEO SUPER-RESOLUTION

Video super-resolution (VSR) methods are commonly classified into traditional and deep learning-based approaches. Schultz and Stevenson [34] introduced a conventional method employing affine models for motion estimation, while 3D steering kernel regression was applied in [35]. Ma et al. [36]utilized the expectation-maximization technique to reconstruct high-resolution frames by estimating the blur kernel. Furthermore, a method in [37] concurrently estimates the blur kernel, motion, and noise level through a Bayesian approach for reconstructing high-resolution frames in VSR.

In recent times, deep learning-based strategies have emerged to address the image super-resolution (SR) challenge. Given that a video comprises a sequence of moving images over time, image SR methodologies can be adapted for VSR by incorporating necessary modifications. Notable deep learning-based image SR models include SRCNN [5], SRGAN [20], FSRCNN [18], ESPCN [6], and ZSSR [14].

VSRnet [21], a model derived from SRCNN, is proposed for video super-resolution.

Deep-learning-based VSR methodologies are commonly categorized into two distinct categories: those incorporating frame alignment and those operating without it. The former class leverages motion estimation and compensation techniques as initial processing steps to extract precise inter-frame motion details [38]–[41] and facilitate frame alignment [10], [42]–[45]. This approach, demonstrated in studies such as [46]–[49], proves particularly effective in scenarios involving significant motion dynamics. The optical flow method that uses variations and correlations in the temporal domain to compute the motion between two nearby frames is popularly used in most motion estimation techniques [50], [51]. The motion compensation methods can also be applied using either the traditional methods [52], [45] or using deep learning-based approaches [51], [50], [53]. Deformable Convolution Methods for video SR were first proposed by [54], [55], [56] [57], [58], [59]. BasicVSR [60] used a simple RNN architecture with propagation, alignment, and upsampling modules to make VSR suitable for real-time applications. Later, for better handling of misalignment, enhanced propagation along with flow-guided deformable alignment is introduced in BasicVSR [60] to the proposed BasicVSR++ [61].

Deformable attention mechanisms have gained prominence in the area of blind video super-resolution (VSR) due to their ability to handle complex spatial transformations and focus on relevant features. One notable work in this domain is the Deep Blind Super-Resolution for Satellite Video [62]. The proposed BSVSR algorithm in [62] is an empirical approach for blind SVSR that emphasizes sharper cues by considering pixel-wise blur levels using the approach called coarse-to-fine manner. It utilizes multi-scale deformable convolution for aggregating the temporal redundancy across adjacent frames through window-slid progressive fusion, followed by deformable attention for meticulous integration of adjacent features into mid-feature.

Another significant contribution is the Bidirectional Multi-scale Deformable Attention for Video Super-Resolution [63]. This method uses a Deformable Alignment Module (DAM) which contains two types of modules: Multi-scale Deformable Convolution Module (MDCM) used to improve the robustness of the adjacent frame alignment process by using the offset information in the different scales and aligning the frames at the feature level. The second module is the Multi-scale Attention Module (MAM), used to extract the local as well as global features of the aligned features.

Moreover, for lightweight VSR [64], the Deformable Spatial-Temporal Attention aggregates the spatial-temporal information obtained from the multiple reference frames into the current frame to improve the reconstruction effect.

These works demonstrate the effectiveness of deformable attention mechanisms in handling the challenges of alignment and fusion in blind VSR tasks. They provide a robust and effective way to enhance the resolution of video sequences while maintaining temporal consistency.

In methods without alignment techniques, alignment is not performed by aligning neighbouring frames; instead, the Spatio-temporal or spatial information is used for executing feature extraction. This technique can further be classified into four types i.e. 2D convolution methods (2D Conv) [65] , non-local network-based methods [59], [66], [67], 3D convolution methods (3D Conv) [68], and Recurrent CNN (R-CNN) based methods [69]. The 2D convolution methods (2D Conv) fall under the umbrella of spatial methods whereas the remaining three methods belong to the Spatio-temporal category and utilize both the temporal as well as spatial information obtained from the input videos [4]. In the 2D convolution method [65], [70], a 2D convolutional network absorbs the correlation information existing within the frames by itself. The three important stages, feature extraction, fusion and SR are performed spatially on the frames passed as input to the model instead of performing actions like motion estimation and motion compensation [4] whereas 3D Convolution Methods [68], [71], [72] utilizes spatial as well as the temporal information to super-resolve the video and Recurrent CNNs [69], [73], [74], [75] can accurately represent the temporal dependency in sequential data and has been used for video SR. The use of direct 3D convolution is introduced by Li et al. [76] to generate video sequences. This eliminates the need for RNNs and allows for efficient processing. To avoid the use of explicit motion compensation, Jo et al. [71] proposed an architecture that learns the dynamic upsampling filters for each pixel based on its local spatio-temporal neighbourhood. CycMu-Net [77] used the cycle-projected mutual learning between spatial and temporal super-resolution tasks to produce optimal results in terms of both detail and consistency. In this method, spatial features refine temporal predictions, while temporal information helps extract finer spatial details. Addressing the issue of blind VSR, DynaVSR [78]utilizes a dynamic encoder-decoder architecture that adapts to different degradation types at runtime. Apart from CNN architecture, transformer architecture is also exploited for VSR. Liang et al. [79] proposed a Recurrent Video Restoration Transformer (RVRT) with guided deformable attention to handle the complex temporal dependencies and object deformations. In recent years, the video inbetweening technique is also utilized aiming to enhance the temporal resolution of video sequences by creating new frames between known keyframes. Initial methods for video inbetweening include optical flow-based interpolation [80], [81] and pixel motion transformation [82], [83]. To perform long-term video interpolation, block-based motion estimation/compensation methods or LSTM models [84] were used.

C. ZERO-SHOT SUPER-RESOLUTION WITH META-LEARNING

The SOTA SISR methods fail, and their performance deteriorates in the case of real LR input (LR image with compression, sensor, and random noises) as they have trained on datasets where LR counterpart is generated synthetically

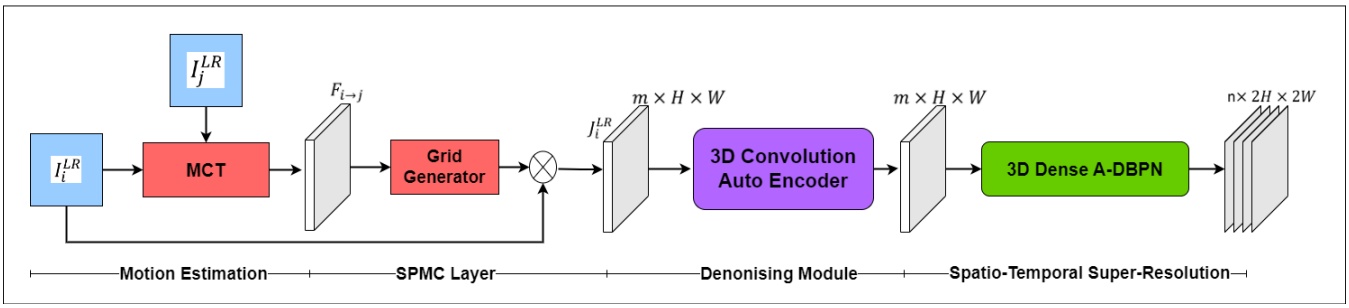

FIGURE 2: Block diagram for the proposed framework. Here H and W are the height and width of the image frame respectively whereas n and m are the number of frames where $m < n$.

by bicubic down-sampling of corresponding HR Image. These SOTA SR methods are very deep and computationally complex, too. To overcome these challenges, Shocher et al. [14] have proposed zero-shot learning for SISR. The model proposed was very light and did not need any training on the external dataset. The proposed method performs image-specific training at test time and outperforms SOTA in the case of real LR images like old photos and degraded images. However, due to its limitation of fast adaptation to datasets having statistically distant noises and kernels, Soh et al. have presented meta-transfer learning along with ZSSR [85] to initialize the weights of the model to take advantage of learning from the external dataset with internal learning in the zero-shot setting. Here, meta-transfer learning helps in fast convergence during training in a zero-shot setting at test time. Mohammad Emad has presented dual-path zero-shot learning [86], which attempts to train cycle GAN-based architecture in the zero-shot setting to improve performance for real-world LR images further. [87] also employs a similar approach to meta-transfer learning.

Performing Spatio-temporal SR, i.e., creating intermittent frames along with SR, is more challenging and intimidating. Spatio-temporal VSR methods [9], [88], [89], [90], primarily work on bicubic and tricubic methods to form LR images, they require large datasets to generalize, they fail to adapt to blurriness created from different kernels and were computationally complex. This limits their use in practical applications such as HD television [91], UAV surveillance [92], [93], security [94], [95], etc. Due to the heterogeneous nature of noise in real LR, these models do not produce effective motion-consistent HR videos.

D. MOTION ESTIMATION AND COMPENSATION

In video super-resolution, simply applying image super-resolution methods to each frame independently may not produce satisfactory results, because there may be inter-frame motion that causes temporal distortion and blur. Inter-frame motion is the movement of objects or cameras between consecutive frames, which can create misalignment and inconsistency between the frames. Therefore, motion estimation and compensation are needed to handle the inter-frame motion and align the frames before applying super-resolution.

To solve the motion estimation problem several deep learning-based networks were used. For stereo matching, Zbontar and LeCun [96] and Lou et al. [97] used to learn patch distance measures by employing a CNN whereas Fischer et al. [98] and Mayer et al [99] proposed the use of end-to-end architectures to predict the optical flow and it's stereo disparity. For the motion compensation problem, earlier VSR methods [100]–[104] achieved the inter-frame motion compensation by either estimating optical flow or by applying block-matching. In deep-learning-based VSR methods use backward warping by aligning all other frames to the reference frame to achieve inter-frame motion compensation.

The existing work on zero-shot super-resolution with meta-learning was limited to images. This motivated us to design an architecture for meta-learning-based zero-shot video space-time SR. Our model incorporates all benefits obtained by the zero-shot and meta-learning-based training. Along with this, our model solves denoising and super-resolution simultaneously with a dedicated architecture for joint optimization of denoising and super-resolution.

III. METHODOLOGY

A. PROPOSED ARCHITECTURE AND OPTIMISATION

The block diagram for the proposed methodology is depicted in Figure 2. The Motion estimation and compensation modules are employed to maintain the temporal coherency in the output video sequence. After the frame alignment, the enhancement module which consists of two parts, i.e., the denoising module and the spatio-temporal super-resolution module, is employed to get spatio-temporal super-resolution. The details of each module are as follows:

1) Motion Estimation

Motion estimation (ME) is the process of estimating the motion vectors between successive frames in a given video sequence. Motion vectors are the displacement of the pixels from one frame to the other, which indicate the direction as well as the magnitude of motion. The ME module takes two LR frames as input and generates an LR motion field as output, and it can be defined as

$$F_{i\to j} = Net_{ME}(I_i^L, I_j^L; \theta_{ME}) \quad (1)$$

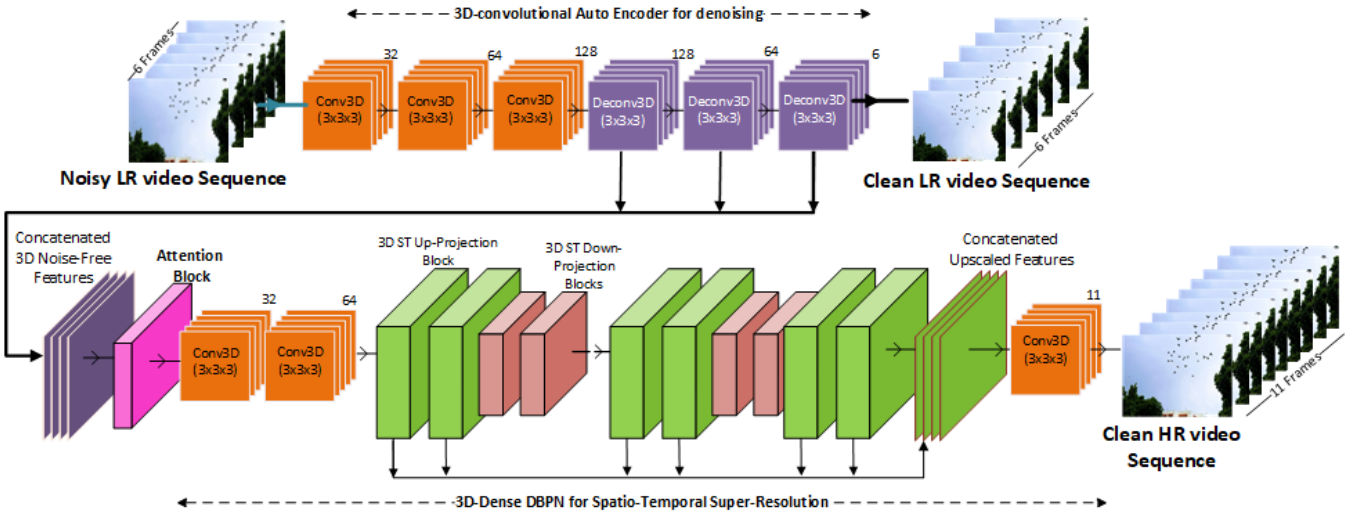

FIGURE 3: Proposed Architecture of 3D-CAE guided Deep Spatio-Temporal Back Projection Network for Video Super-Resolution

where $F_{i\to j} = (u_{i\to j}, v_{i\to j})$ is the motion field calculted from frame I_i^L, I_j^L, and θ_{ME} is the set of motion estimation parameters. The use of neural networks for motion estimation is a well-established concept, and several existing works [98]–[100], [105] have demonstrated significant achievements in this area. In our study, we tested FlowNet-S [98] and the motion compensation transformer (MCT) module from VESPCN [100] and opted for MCT due to its lower parameter count and, consequently, reduced computational cost. The MCT uses self-attention along with cross-attention mechanisms to capture the long-range dependencies and temporal alignment between the frames. The self-attention mechanism computes the similarity between each pixel and all other pixels within the same frame and generates a weighted sum of the pixel features. The cross-attention mechanism calculates the similarity between each pixel in the reference frame and all other pixels in the neighbouring frames and generates a weighted sum of the pixel features. The output of the MCT is a set of motion vectors for each pixel in the reference frame.

2) Motion Compensation

Motion compensation is the process of aligning the frames according to the motion vectors to reduce temporal distortion and blur. Motion compensation can help to improve the temporal consistency and details of the video. We used the motion compensation layer proposed by Tao et al. [43] that leverages sub-pixel information from motion to achieve simultaneous sub-pixel motion compensation (SPMC) The SPMC performs sub-pixel shifting and interpolation to generate aligned frames. The sub-pixel shifting operation shifts each pixel in the neighboring frames according to its corresponding motion vector, and produces a set of shifted frames. The interpolation operation combines the shifted frames using a convolutional filter and produces an aligned frame. The output of the SPMC is a set of aligned frames that match the reference frame. It can be defined as

$$J_i^{LR} = L_{SPMC}(I_i^{LR}, F; \alpha) \tag{2}$$

where I_i^{LR} is the LR image and J_i^{LR} is the output image, F denotes the optical flow utilized for transposed warping, and α represents the scaling factor. As illustrated in Figure 2, transformed coordinates are first calculated using the estimated motion flow $F = (u, v)$ and can be expressed as:

$$\begin{pmatrix} x_p^s \\ y_p^s \end{pmatrix} = W_{F;\alpha} \begin{pmatrix} x_p \\ y_p \end{pmatrix} = \alpha \begin{pmatrix} x_p + u_p \\ y_p + v_p \end{pmatrix} \tag{3}$$

here, p indexes pixels in the LR image space, where x_p and $y_p$represent the two coordinates of p. Additionally, u_p and v_p denote the flow vectors estimated from the previous stage. We use the operator $W_{F;\alpha}$ to denote the transformation of coordinates, which depends on the flow field F and the scale factorα. Subsequently, x_p^s and y_p^s refer to the transformed coordinates in output image space. Finally, the resulting image J_q^H can be constructed in the output image space using

$$J_q^H = \sum_{p=1} J_p^L M(x_p^s - x_q) M(y_p^s - y_q) \tag{4}$$

where q indexes output image pixels, x_q and y_q are the coordinates for pixel q in the output image grid. $M(\cdot)$ is the sampling kernel.

3) The Enhancement Module

The detailed architecture for the enhancement module is shown in Figure 3, which is used to improve the spatial and temporal resolution of real LR video. Since the input video is of low resolution and has noise and degradation, improving the resolution of the given input real LR video along with removing the noise embedded in it is the main objective of the proposed enhancement module.

TABLE 1: Performance comparison of various video SR algorithms on Vid4 [44] dataset with different densities of noise. Here Rn denotes random noise. The top 2 results are shown in bold

Methods	Vid4 (PSNR/SSIM)	Vid4 + Gaussian (0.01) (PSNR/SSIM)	Vid4 + Gaussian (0.1) (PSNR/SSIM)	Vid4 + Gaussian + Rn (PSNR/SSIM)
TechnoGAN [106]	25.89/-	21.68/0.6482	20.73/0.6341	17.49/0.5934
MMCNN [107]	26.28/0.7844	24.19/0.7194	23.37/0.7059	20.78/0.6493
MEMC-Net [108]	24.37/0.8380	21.72/0.7938	20.18/0.7614	18.94/0.7214
MuCAN [109]	**30.88/0.8750**	**27.04/0.8137**	**25.17/0.8046**	**21.31/0.7419**
DUF [71]	27.38/0.8329	23.28/0.6400	22.76/0.6233	18.88/0.5030
FSTRN [90]	29.95/0.8700	22.98/0.6132	19.72/0.5673	16.39/0.4478
EDVR [110]	27.85/0.8503	24.11/0.6897	23.02/0.6338	19.87/0.5348
TDAN [111]	26.58/0.8010	23.78/0.6429	21.49/0.6176	19.46/0.5937
BasicVSR [60]	27.24/0.8251	26.41/0.8137	23.57/0.7785	21.81/0.7400
BasicVSR++ [61]	27.79/0.8400	26.53/0.8157	24.16/0.7800	22.56/0.7417
RVRT [79]	27.99/0.8462	26.84/0.8162	24.57/0.7846	22.94/0.7468
Ada-VSR [87]	26.98/0.8400	24.83/0.8251	23.18/0.8143	21.48/0.7804
Proposed	**30.07/0.8643**	**27.89/0.8017**	**25.88/0.7916**	**23.16/0.7649**

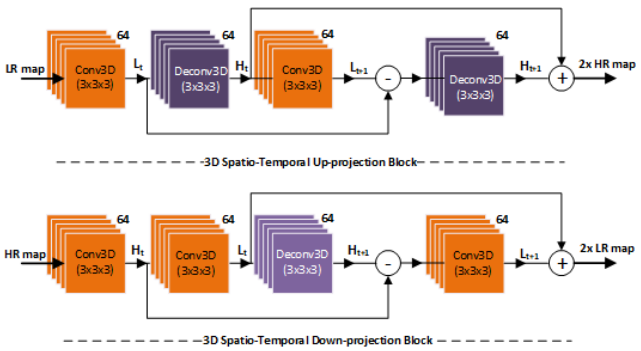

FIGURE 4: Internal Architecture of 3D Spatio-Temporal Up and Down Projection Blocks

The enhancement module consists of two sub-modules; one is for denoising, and another is for spatio-temporal super-resolution. The denoising module is used to remove the unwanted noise and degradation in real LR video. It is a 3D-deep convolutional auto-encoder (3D-CAE) that consists of three 3D-convolution layers followed by three 3D-convolutional transpose layers. The output of this 3D-CAE will be noise-free LR video with the same level of spatial and temporal resolution as in the input.

The features from 3D-deconvolution layers are concatenated as feed to the Spatio-temporal super-resolution module, which is an ADST-BPN. In the spatio-temporal super-resolution module, a 3D-attention layer is present at the beginning, followed by one 3D-convolution layer. After that, there are a series of spatio-temporal up-projection and down-projection blocks arranged in a cascade manner. These spatio-temporal blocks, as shown in Figure 4, are discussed in detail in the next section. This spatio-temporal up projection and down projection is the 3D version of the up and down projection block given by [112]. All the features from Spatio-temporal up-projection blocks are concatenated at last, and these concatenated features are then passed through one more 3D-convolutional layer. The output of this final 3D-convolutional layer is our desired HR output video. In the enhancement module, one 3D-self-attention layer is also used to learn the relation of prominent features that contribute mostly to feature enhancement.

4) Spatial-temporal up and back projection

As given in figure 4, the spatio-temporal up-projection unit takes LR video features coming from previous calculations and maps it to intermediate HR features H_t (spatio-temporal upscaled) using a 3D-convolution and 3D deconvolutional layer employed in series. First, the low-resolution feature at time t is upscaled using a 3D-deconvolution layer to get the high-resolution feature H_t. Mathematically this up-projection can be defined as:

$$H_t = (L_t * p_t) \uparrow_s \tag{5}$$

where s is the scale factor of upsampling and downsampling and p_t is the 3D-deconvolution layer. we have selected $s = 2$ for our proposed framework. Then one more 3D-convolutional g_t layer maps H_t to intermediate LR feature L_{t+1}. This LR feature can be written as:

$$L_{t+1} = (H_t * g_t) \downarrow_s \tag{6}$$

Then we calculate the residual e_t between this LR feature map L_{t+1} and first calculated LR map L_t coming from the first 3D-convolutional layer, defined as:

$$r_t = L_{t+1} - L_t \tag{7}$$

This residue is again passed through a 3D-deconvolutional layer q_{t+1} to get one more intermediate HR map H_{t+1}.

$$H_{t+1} = (r_t * q_{t+1}) \uparrow_s \tag{8}$$

The final Spatio-temporal upscaled HR feature map is achieved by adding outputs of two 3D-deconvolutional layers.

$$HR_{map} = H_{t+1} + H_t \tag{9}$$

The spatio-temporal down-projection unit, as shown in figure 4, works similarly to the spatio-temporal up-projection layer. It takes previously computed spatio-temporal up-scaled HR features as input and produces the final LR using a 3D-convolution layer M_t feature map defined as:

$$L_t = (H_t * M_t) \downarrow_s \tag{10}$$

Then, L_t is upsampled using 3D-deconvolution layer N_t and it can be written as:

$$H_{t+1} = (L_t * N_t) \uparrow_s \tag{11}$$

Now, similar to the Up-projection block, residue r_n is calculated by subtracting the HR feature $Ht + 1$ with intermediate HR feature H_t,

$$r_n = H_{t+1} - H_t \tag{12}$$

Next, we downscale this calculated residual feature using a 3D-convolution layer R_{t+1} to get the LR feature defined as:

$$L_{t+1} = (r_n * R_{t+1}) \downarrow_s \tag{13}$$

Final LR features are obtained by adding the intermediate LR features maps L_t and L_{t+1}.

$$LR_{map} = L_t + L_{t+1} \tag{14}$$

B. LEARNING ON EXTERNAL DATASET

Assume the high-quality dataset D_{HR} having n number of pairs of HR video and corresponding LR video (V_{HR}, V_{LR}), Here V_{LR} is synthetically generated by bicubic downsampling of HR Video frames and dropping of mid frames. Then noise is added to the LR video frames to generate noisy or degraded LR video V_{NLR}. Now our proposed model (ZS-RW-ZSSR) given in Figure 3, is learning the robust spatio-temporal SR mapping f_θ, where θ is the parameter of the model, between noisy-LR and HR Video pairs (V_{NLR}, V_{HR}) and also learning the LR video denoising mapping $f_{\theta CAE}$ by 3D-CAE from noisy-LR and LR video pairs (V_{NLR}, V_{LR}) by minimizing the loss

$$\begin{aligned} L^D(\theta) = w_1 \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \parallel V_{HRi} - f_\theta(V_{NLRi}) \parallel_2 + \\ w_2 \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \parallel V_{LRi} - f_{\theta CAE}(V_{NLRi}) \parallel_2 \end{aligned} \tag{15}$$

here, n is the total number of pairs of LR-HR videos in the external dataset. minimization of weighted linear combination of denoising loss $loss_{denoising}$ and super-resolution loss $loss_{SR}$ $w_1 loss_{SR}$ + $w_2 loss_{denoising}$ is performed to update the weights of model. here, to optimise the weights of ADST-BPN $w_1 = 1$ and $w_2 = 0$ and to optimize the weights of 3D-CAE, $w_1 = 0.3$ and $w_2 = 0.7$. w_1 and w_2 have been decided by trial and error. 3D-CAE and ADST-BPN are end-to-end optimized.

C. META TRANSFER LEARNING FOR VIDEO SR

After learning model parameters (θ) on an external dataset, the model is adapted to different settings (different down-scaling kernels + random/Gaussian noise/blur). One another synthetic dataset has been generated for meta-transfer learning, denoted as D_{Meta}. D_{Meta} consist of pairs (V_{HR}, ${V_{LR}}^k$), here k represents different down-scaling kernel and degradation settings.

The Meta dataset is further divided into task-level training and testing datasets ($D_{training}$ and $D_{testing}$).

For each new task T_i, parameters are updated using one or more gradient decent updates. The new parameter θ_i is then

$$\theta_i = \theta - \alpha \nabla \theta L_{T_i}^{training}(\theta) \tag{16}$$

here, α is a task level learning rate, θ is old parameter.

The model parameters are further optimized by minimizing the loss,

$$\theta_M = argmin_\theta \sum_{T_i} L_{T_i}^{test} \left(\theta - \alpha \nabla_\theta L_{T_i}^{training}(\theta) \right) \tag{17}$$

The model parameters θ_M are generated by the above optimization equation

D. ADAPTION TO TEST IMAGE: META TEST

Then θ_M (pre-trained weight) is used as initial weights to train our proposed enhancement module as given in Figure 3 in a zero-shot manner [14], [85] as shown in Figure 1.

This step is one test video-specific training at test time. Given a test LR video sequence V_{LR}, first, we will down-sample each frame spatially and drop the even frames to generate its spatio-temporal son V_{son}. We performed some epoch of gradient update with V_{son} as input and V_{LR} as a ground truth. After a few epoch updates, we feed the given test sequence V_{LR} as input to the updated learned model to spatio-temporally super-resolve the test video sequence.

IV. EXPERIMENTAL RESULTS

In this section, we exhibit the implementation details of our proposed method. Also, the outcomes of our proposed network and its comparison with existing SOTA methods for video SR. The comparative analysis of our experiment in terms of evaluation metrics, as well as the visual comparison, is also presented. This section also discusses the ablation study and information about computational complexity.

A. IMPLEMENTATION DETAILS

Dataset: We used two popular video super-resolution datasets for training and evaluation of our proposed model: Vid4 [44], and REDS [113]. The **Vid4** dataset is a commonly used benchmark for video super-resolution algorithms. It consists of four video sequences: “walk,” “city,” “foliage,”

TABLE 2: Performance comparison of various video SR algorithms on REDS4 [113] dataset with different densities of noise. Here Rn denotes random noise. The top 2 results are shown in bold

Methods	REDS4 (PSNR/SSIM)	REDS4 + Gaussian (0.01) (PSNR/SSIM)	REDS4 + Gaussian 0.1 (PSNR/SSIM)	REDS4 + Gaussian + Rn (PSNR/SSIM)
TechnoGAN [106]	28.49/0.7816	26.16/0.7613	25.31/0.7543	21.09/0.7201
MMCNN [107]	29.17/0.7904	27.44/0.7643	26.07/0.7540	22.51/0.7289
MEMC-Net [108]	28.49/0.8507	27.01/0.8312	26.18/0.8176	23.76/0.7819
MuCAN [109]	**31.46/0.8934**	**28.41/0.8629**	**27.84/0.8537**	**24.12/0.8196**
DUF [71]	28.63/0.8056	25.63/0.8881	23.43/0.6939	20.05/0.6008
FSTRN [90]	27.90/8634	25.52/0.7658	23.19/0.6431	18.35/0.5876
EDVR [110]	28.88/0.8361	26.34/0.8300	24.37/0.7466	21.00/0.6134
TDAN [111]	29.71/0.8214	27.09/0.7943	26.17/0.7649	22.96/0.7319
BasicVSR [60]	31.42/0.8909	29.83/0.8713	8.53/0.8571	24.83/0.8017
BasicVSR++ [61]	32.39/0.9069	30.18/0.8776	28.73/0.8618	25.12/0.8134
RVRT [79]	32.75/0.9113	31.43/0.8819	29.87/0.8632	26.43/0.8193
Proposed	**31.05/0.8846**	**29.87/0.8716**	**29.18/0.8619**	**26.49/0.8278**

TABLE 3: Performance comparison of our proposed CAST-Net with existing methods on RealVSR and MVSR4x dataset.

Methods	RealVSR [114]	MVSR4× [115]
	SNR/SSIM/LPIPS	PSNR/SSIM/LPIPS
TDAN	23.71/0.7737/0.229	23.07/0.7492/0.282
EDVR	23.96/0.7781/0.216	23.51/0.7611/0.268
BasicVSR	24.00/0.7801/0.209	23.38/0.757594/0.270
BasicVSR++	24.24/0.7933/0.216	23.70/0.7713/0.263
EAVSR+	24.41/0.7953/0.212	23.94/0.7726/0.259
CASTNet (Ours)	**25.13/0.8061/0.207**	**24.61/0.7926/0.247**

TABLE 4: Estimated FLOPS for each component of our proposed model.

Component	Estimated FLOPSc(G)
CNN Blocks	103.12
PReLU	2.9
Addition	2.6
Subtraction	2.6
Concatenation	2.2
Attention	2.2
Total	115.6

and "calendar." The video's length ranges from 26 to 47 frames, and the resolution of videos ranges from 704x576 to 740x480. The **REDS (Realistic and Dynamic Scenes)** dataset is a large-scale video super-resolution dataset. It is designed for video deblurring and video SR tasks. It utilises 120 fps videos to create blurry frames through the merging of consecutive frames.

These datasets provide a diverse range of video sequences, which are essential for training robust and generalizable video super-resolution models. The high-quality ground truth frames in these datasets allow for precise evaluation of the super-resolution performance.

Degraded Video Sequence Generation: To add realistic degradation in images for video super-resolution, we used the degradation model that simulates real-world degradation. Several models [24], [25], [116], [117] have been proposed that can be used for this purpose. By using these models, we generated degraded LR images that resemble real-world degradation and used them to train the proposed VSR model.

The proposed model was initially trained on a large-scale dataset for 200 epochs, utilizing the Adam optimizer having a learning rate of 0.0001. **Meta-learning Configuration**: Our meta-learning configuration is based on the MAML (Model-Agnostic Meta-Learning) framework. For meta-transfer learning, we used the Model-Agnostic Meta-Learning (MAML) algorithm with an inner learning rate of 0.01 and an outer learning rate of 0.001. The meta-training was performed for 50 epochs. For Motion Compensation Transformer (MCT), a pretrained MCT model is used that is trained on a large-scale video dataset for motion estimation. The MCT module underwent the fine-tuning process during the training of the entire model to adapt the specific characteristics of the degraded and noisy real low-resolution videos.

TABLE 5: Ablation study to show the importance of Meta-Learning, denoising module and attention layer in Proposed Video 4X Real VSR architecture on Vid4 [44]+G 0.01 and REDS4 [113]+G 0.1

Algorithms\Datasets	REDS4+G 0.1 PSNR/SSIM	Vid4+G 0.01 PSNR/SSIM
Proposed	**25.89/0.7889**	**24.34/0.7123**
Proposed without Meta Learning	24.32/0.7028	21.39/0.6882
Proposed without Meta Learning and Denoising module	23.76/0.6823	21.11/0.6543
Proposed without Meta learning, Denoising and attention module	23.65/0.6519	21.04/0.6329

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406476

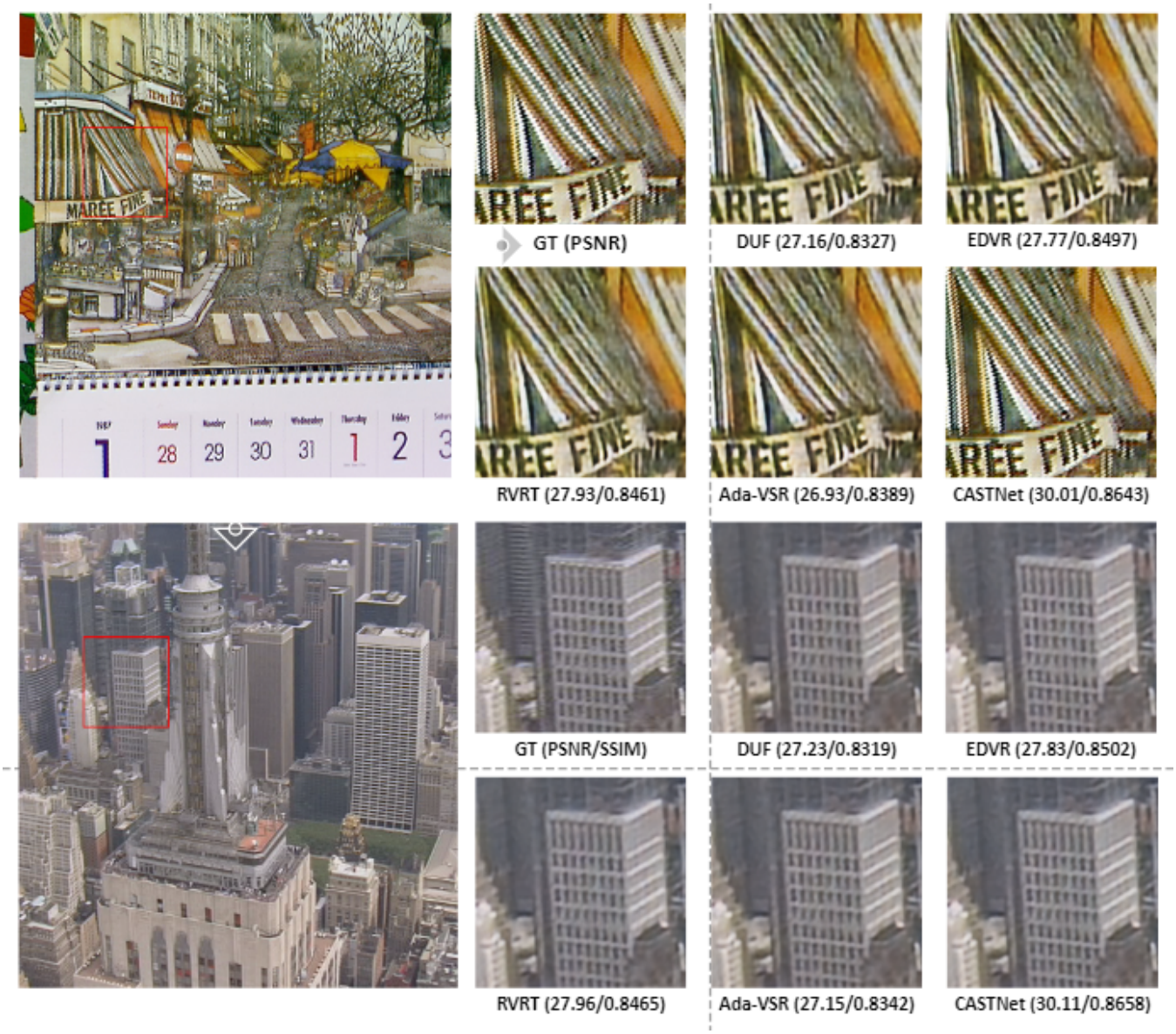

FIGURE 5: Qualitative Comparison of Spatial SR on Vid4 Dataset.

B. RESULT ANALYSIS

The research work is evaluated in two ways: first, quantitative evaluation, and second, qualitative evaluation. In quantitative evaluation, we compare the achieved result using PSNR and SSIM metrics with other SOTA methods for video super-resolution. Gaussian noise is added with different densities in test images for evaluation also in one case Random noise is added in the test images as it is evident from Table 1 and 2, PSNR/SSIM values are closer to MuCAN [109] when no noise is present, but PSNR/SSIM increases compared to other methods as we include the noise. This is because the proposed model is specifically trained on real-world noisy video frames. The proposed method surpasses the other methods for all varieties of noise densities and on both datasets.

In qualitative evaluation, the visual results of the proposed methods are presented and compared to the visual results of other SOTA methods. In Figure 5, a comparison of the proposed CASTNet with SOTA methods is shown, whereas in Figure 6 and 7, a comparison of spatial super-resolution on REDS dataset is shown. In all these figures, it can be observed that the proposed method can reconstruct much finer details and texture. Also, in the temporal domain, mid-frames between two successive frames of the given input video sequence are being constructed. The proposed model is also able to reconstruct the mid-frames with less amount of flicker. This is due to the better reconstruction of sharpness and smooth edges in mid-frames. In Figure 8, visual results of temporal super-resolution are shown for two types of video sequences. For both video sequences, row (a) is the original video frame sequence, and row (b) is the reconstructed video

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406476

FIGURE 6: Qualitative Comparison of Spatial SR on REDS Dataset.

frames, which show the better transition from one frame to the next frame.

We conducted an additional experiment where we quantitatively measured the flicker by calculating the temporal intensity variation between consecutive frames. This provides an objective measure of the flicker. This quantitative analysis, provided in Table 7, indicates that our proposed methods generated mid-frames almost similar to the ground truth.

In Figure 9, a comparison of the temporal super-resolution performance of the proposed model is shown with the traditional interpolation approach. In any video sequence, when the temporal frequency of a moving object is greater than the camera frame rate, then the aliasing problem occurs. This aliasing problem is known as Motion aliasing or Temporal aliasing. In this aliasing problem, the object seems to move to a false trajectory, or its motion could be distorted. This problem can be solved by reconstruction of mid-frames between successive frames. In figure 9, the real movement direction is clockwise, but on the left side, the fan seems to be rotating in another direction. On the right side, this aliasing effect is reduced by using the reconstructed mid-frames.

C. COMPUTATIONAL COMPLEXITY

This study presents an extensive investigation into the computational complexity of our proposed model, focusing primarily on the number of Floating Point Operations Per

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406476

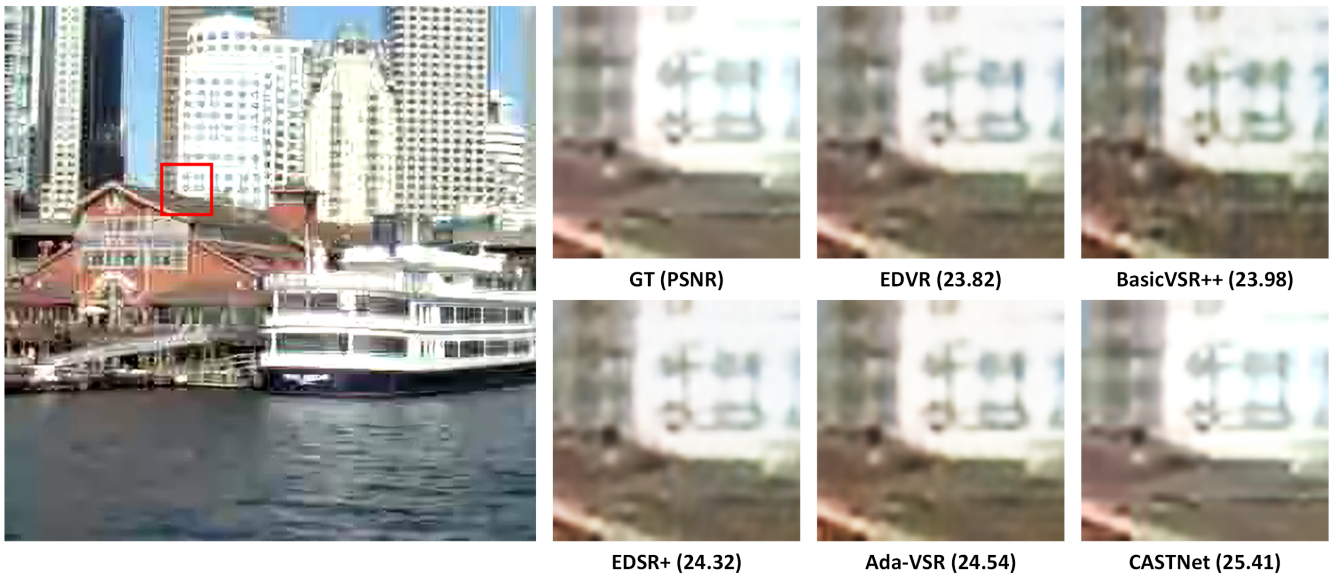

FIGURE 7: Qualitative Comparison of Spatial Video on RealVSR [114] Dataset.

FIGURE 8: Visual Results of Temporal SR on Vid4 Dataset. (a) and (c) Original Frames (b) and (d) Predicted Frames. Even numbered frames are reconstructed frames between two successive odd-numbered frames.

TABLE 6: Comparison of model efficiency of our proposed method with existing methods on the basis of Model size, runtime, and memory for LR input of 320×180

Method	#Params (M)	Runtime (ms)	Memory (M)
BasicVSR++	7.3	77	223
EDVR	20.6	378	3535
VSRT	32.6	328	27487
VRT	10.8	183	1056
DUF	5.8	974	-
Proposed	1.69	56	6.48

Second (FLOPs). Our model demonstrates exceptional performance in various applications, achieving state-of-the-art results while maintaining a balance between computational efficiency and performance. Our model consists of six main components: 3D Convolutional Neural Network (3D-CNN) blocks, PReLU, addition, subtraction, concatenation, and attention. Each component contributes to the total computational load of the model, making it imperative to analyze each part individually. Table 4 summarizes the FLOPs breakdown for all major components of our model.

Additionally, We demonstrate the comparison of our proposed model with existing models in terms of model size, runtime, and memory in Table 6. Our proposed model stands out with only 1.69 million parameters, demonstrating efficient parameter utilization. It boasts a fast runtime of 56 ms

FIGURE 9: Comparison of Temporal Super-Resolution (3x): Bicubic Interpolation Method (Left) vs. Our Method (Right)

TABLE 7: Quantitative analysis of temporal intensity variation for Vid4 dataset.

Category	City	Walk	Foliage	Calender
Groun Truth	97.53	113.50	125.29	153.13
Generated	96.82	112.67	124.47	152.91

and minimal memory consumption at 6.48 MB, showcasing superior computational efficiency without compromising performance. Our proposed model excels in performance metrics by demonstrating efficient parameter utilization, fast runtime, and minimal memory consumption compared to existing models.

D. ABLATION STUDY

In Table 5, it can be seen that, with meta-learning, we see a 1.5 dB improvement in PSNR as it harnesses the benefits of pre-trained weights. It also helped in obtaining kernel-agnostic properties for the model. We can also see, that without the denoising module, the PSNR decreases significantly. The denoising module improved the quality of the LR image, which is fed to our enhancement (ADST-BPN) module, which significantly improved the quality of the SR Video sequence. The attention module helped in localizing the optical flow in the video which significantly helped in recovering and improving the temporal and motion coherence in the video.

V. LIMITATIONS AND FUTURE WORK

Limitations: The proposed framework may not be as effective in scenarios with extremely low-resolution videos or highly noisy environments. The proposed framework may not be as effective in scenarios with complex motion patterns or large motion displacements. The proposed framework may not be as effective in scenarios with limited training data or when the training data does not accurately represent the target domain. The proposed framework may not be as effective in scenarios with limited computational resources or when real-time performance is required.

Future points: Extending the proposed framework to handle more complex scenarios, such as extremely low-resolution videos, multi-modal data, and real-time applications. Investigating the potential of the proposed framework for other computer vision tasks, such as video denoising, video deblurring, and video inpainting. Investigating the application of the suggested framework in different areas, including medical imaging, remote sensing, and security surveillance. To further validate the effectiveness of the proposed framework, it is important to assess its performance on more extensive and varied datasets. Additionally, employing alternative evaluation metrics like the Perceptual Index (PI) could offer deeper insights.

VI. CONCLUSION

This paper presents a zero-shot learning and meta-learning-based video space-time Super-Resolution (SR) algorithm. A novel noise-robust video space-time SR architecture, namely, 3D-Deep Convolutional Auto-Encoder guided attention-based deep spatio-temporal back-projection network (CAST-Net), is introduced. This proposed method can effectively handle real degradation and noises while super-resolving Low-Resolution (LR) video by jointly optimizing two different SR and denoising losses. The proposed solution converges faster and is robust against realistic degradation. Several comparative studies are conducted and shown in the experiment section, to validate the efficacy of the proposed framework. A detailed ablation study is also conducted to highlight the significance of each component in the methodology. We also proposed some limitations and future points of our proposed work. One possible limitation is that our proposed framework may not be as effective in scenarios with extremely low-resolution videos, highly noisy videos, complex motion pattern scenarios, or scenarios with large motion displacements. Looking forward, we plan to address these limitations in our future work. We aim to extend our framework for other computer vision tasks, for example, video denoising and video inpainting. We also plan to explore the possibility of using our framework for other domains, such as remote sensing and medical imaging. We also plan to test our model on a wider variety of videos to ensure its robustness and generalizability. We believe that our work provides a solid foundation for future research in video space-time SR and opens up new avenues for exploration.

REFERENCES

[1] Y. Xiao, Q. Yuan, K. Jiang, X. Jin, J. He, L. Zhang, and C.-w. Lin, "Local-global temporal difference learning for satellite video super-resolution," IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2023.

[2] Y. Xiao, Q. Yuan, J. He, Q. Zhang, J. Sun, X. Su, J. Wu, and L. Zhang, "Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer," International Journal of Applied Earth Observation and Geoinformation, vol. 108, p. 102731, 2022.

[3] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–19, 2021.

[4] H. Liu, Z. Ruan, P. Zhao, F. Shang, L. Yang, and Y. Liu, "Video super resolution based on deep learning: A comprehensive survey," arXiv preprint arXiv:2007.12928, 2020.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in European conference on computer vision. Springer, 2014, pp. 184–199.

[6] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.

[8] M. Sharma, R. Mukhopadhyay, A. Upadhyay, S. Koundinya, A. Shukla, and S. Chaudhury, "Irgun: Improved residue based gradual up-scaling network for single image super resolution," 2018, pp. 834–843.

[9] M. Sharma, S. Chaudhury, and B. Lall, "Space-time super-resolution using deep learning based framework," in Pattern Recognition and Machine Intelligence - 7th International Conference, PReMI 2017, Kolkata, India, December 5-8, 2017, Proceedings, ser. Lecture Notes in Computer Science, vol. 10597. Springer, 2017, pp. 582–590.

[10] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4778–4787.

[11] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Investigating tradeoffs in real-world video super-resolution," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 5962–5971.

[12] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[13] X. Yang, W. Xiang, H. Zeng, and L. Zhang, "Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4781–4790.

[14] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3118–3126.

[15] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution-supplementary material."

[16] R. Sharma, M. Sharma, A. Shukla, and S. Chaudhury, "Conditional deep 3d-convolutional generative adversarial nets for rgb-d generation," Mathematical Problems in Engineering, vol. 2021, pp. 1–8, 2021.

[17] A. Shukla, A. Upadhyay, S. Bhugra, and M. Sharma, "Opinion unaware image quality assessment via adversarial convolutional variational autoencoder," 2024, pp. 2153–2163.

[18] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in European conference on computer vision, 2016, pp. 391–407.

[19] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem, pp. 1646–1654, 2016.

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[21] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," IEEE Transactions on Computational Imaging, vol. 2, no. 2, pp. 109–122, 2016.

[22] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3262–3271.

[23] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 945–952.

[24] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," Advances in Neural Information Processing Systems, vol. 32, 2019.

[25] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1604–1613.

[26] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3867–3876.

[27] P. Behjati, P. Rodriguez, C. F. Tena, A. Mehri, F. X. Roca, S. Ozawa, and J. Gonzalez, "Frequency-based enhancement network for efficient super-resolution," IEEE Access, vol. 10, pp. 57 383–57 397, 2022.

[28] F. Luo, X. Wu, and Y. Guo, "And: Adversarial neural degradation for learning blind image super-resolution," in Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[29] J. Dong, H. Bai, J. Tang, and J. Pan, "Deep unpaired blind image super-resolution using self-supervised learning and exemplar distillation," International Journal of Computer Vision, pp. 1–14, 2023.

[30] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," Information Fusion, vol. 96, pp. 297–311, 2023.

[31] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, "Sinsr: Diffusion-based image super-resolution in a single step," 2023.

[32] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution," IEEE Transactions on Geoscience and Remote Sensing, 2023.

[33] Y. Guo, X. Zhang, and X. Wu, "Deep multi-modality soft-decoding of very low bit-rate face videos," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3947–3955.

[34] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," IEEE transactions on image processing, vol. 5, no. 6, pp. 996–1011, 1996.

[35] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," IEEE Transactions on Image Processing, vol. 18, no. 9, pp. 1958–1975, 2009.

[36] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5224–5232.

[37] C. Liu and D. Sun, "On bayesian adaptive video super resolution," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 2, pp. 346–360, 2013.

[38] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 531–539.

[39] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in European conference on computer vision. Springer, 2004, pp. 25–36.

[40] C. Liu et al., "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[41] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1744–1757, 2011.

[42] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," IEEE Transactions on Computational Imaging, vol. 2, no. 2, pp. 109–122, 2016.

[43] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4472–4480.

[44] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang, "Robust video super-resolution with learned temporal dynamics," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2507–2515.

[45] M. Drulea and S. Nedevschi, "Total variation regularization of local-global optical flow," in 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2011, pp. 318–323.

[46] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6626–6634.

[47] T. H. Kim, M. S. Sajjadi, M. Hirsch, and B. Scholkopf, "Spatio-temporal transformer network for video restoration," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 106–122.

[48] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," International Journal of Computer Vision, vol. 127, no. 8, pp. 1106–1125, 2019.

[49] M. Haris, G. Shakhnarovich, and N. Ukita, "Space-time-aware multi-resolution video enhancement," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2859–2868.

[50] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2758–2766.

[51] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4161–4170.

[52] B. D. Lucas, T. Kanade et al., "An iterative image registration technique with an application to stereo vision." Vancouver, British Columbia, 1981.

[53] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470.

[54] X. Wang, K. Chan, K. Yu, C. Dong, and C. E. Change Loy, "Video restoration with enhanced deformable convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 2019, pp. 16–17.

[55] H. Wang, D. Su, C. Liu, L. Jin, X. Sun, and X. Peng, "Deformable non-local network for video super-resolution," IEEE Access, vol. 7, pp. 177 734–177 744, 2019.

[56] J. Chen, X. Tan, C. Shan, S. Liu, and Z. Chen, "Vesr-net: The winning solution to youku video enhancement and super-resolution challenge," arXiv preprint arXiv:2003.02115, 2020.

[57] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.

[58] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.

[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[60] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4947–4956.

[61] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5972–5981.

[62] Y. Xiao, Q. Yuan, Q. Zhang, and L. Zhang, "Deep blind super-resolution for satellite video," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–16, 2023.

[63] Z. Zhou, B. Xue, H. Wang, and J. Zhao, "Bidirectional multi-scale deformable attention for video super-resolution," Multimedia Tools and Applications, pp. 1–22, 2023.

[64] T. Xue, X. Huang, and D. Li, "Deformable spatial-temporal attention for lightweight video super-resolution," in Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer, 2023, pp. 482–493.

[65] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

[66] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3106–3115.

[67] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "Mucan: Multi-correspondence aggregation network for video super-resolution," in European Conference on Computer Vision. Springer, 2020, pp. 335–351.

[68] X. Jia, B. De Brabandere, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," in NIPS, 2016.

[69] R. Cai, X. Zhang, and H. Wang, "Bidirectional recurrent convolutional neural network for relation classification," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 756–765.

[70] B. Yan, C. Lin, and W. Tan, "Frame and feature-context video super-resolution," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 5597–5604.

[71] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "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, 2018, pp. 3224–3232.

[72] S. Y. Kim, J. Lim, T. Na, and M. Kim, "Video super-resolution based on 3d-cnns with consideration of scene change," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 2831–2835.

[73] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 1015–1028, 2017.

[74] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, and Z. Xue, "Residual invertible spatio-temporal network for video super-resolution," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 5981–5988.

[75] J.-H. Jacobsen, A. Smeulders, and E. Oyallon, "i-revnet: Deep invertible networks," arXiv preprint arXiv:1802.07088, 2018.

[76] Y. Li, D. Roblek, and M. Tagliasacchi, "From here to there: Video inbetweening using direct 3d convolutions," 2019.

[77] M. Hu, K. Jiang, L. Liao, J. Xiao, J. Jiang, and Z. Wang, "Spatial-temporal space hand-in-hand: Spatial-temporal video super-resolution via cycle-projected mutual learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3574–3583.

[78] S. Lee, M. Choi, and K. M. Lee, "Dynavsr: Dynamic adaptive blind video super-resolution," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2093–2102.

[79] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool, "Recurrent video restoration transformer with guided deformable attention," Advances in Neural Information Processing Systems, vol. 35, pp. 378–393, 2022.

[80] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470.

[81] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4463–4471.

[82] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 261–270.

[83] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9000–9008.

[84] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[85] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3516–3525.

[86] M. Emad, M. Peemen, and H. Corporaal, "Dualsr: Zero-shot dual learning for real-world super-resolution," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1630–1639.

[87] A. Gupta, P. Jonnalagedda, B. Bhanu, and A. K. Roy-Chowdhury, "Ada-vsr: Adaptive video super-resolution with meta-learning," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 327–336.

[88] Z. Xiao, Z. Xiong, X. Fu, D. Liu, and Z.-J. Zha, "Space-time video super-resolution using temporal profiles," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, p. 664–672.

[89] O. Shahar, A. Faktor, and M. Irani, "Space-time super-resolution from a single video," in Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, 2011, p. 3353–3360.

[90] B. D. L. Z. Y. X. D. T. Sheng Li, Fengxiang He, "Fast spatio-temporal residual network for video super-resolution," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.

[91] C. C. Bracken, "Presence and image quality: The case of high-definition television," Media psychology, vol. 7, no. 2, pp. 191–205, 2005.

[92] A. Puri, "A survey of unmanned aerial vehicles (uav) for traffic surveillance," Department of computer science and engineering, University of South Florida, pp. 1–29, 2005.

[93] E. Semsch, M. Jakob, D. Pavlicek, and M. Pechoucek, "Autonomous uav surveillance in complex urban environments," in 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 2. IEEE, 2009, pp. 82–85.

[94] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," Journal of Visual Communication and Image Representation, vol. 77, p. 103116, 2021.

[95] N. Haering, P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," Machine Vision and Applications, vol. 19, no. 5-6, pp. 279–290, 2008.

[96] J. Zbontar, Y. LeCun et al., "Stereo matching by training a convolutional neural network to compare image patches." J. Mach. Learn. Res., vol. 17, no. 1, pp. 2287–2318, 2016.

[97] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5695–5703.

[98] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2758–2766.

[99] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4040–4048.

[100] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4778–4787.

[101] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," IEEE transactions on image processing, vol. 13, no. 10, pp. 1327–1344, 2004.

[102] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 531–539.

[103] C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in CVPR 2011. IEEE, 2011, pp. 209–216.

[104] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," IEEE Transactions on Image Processing, vol. 18, no. 9, pp. 1958–1975, 2009.

[105] J. J. Yu, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, 2016, pp. 3–10.

[106] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning temporal coherence via self-supervision for gan-based video generation," ACM Transactions on Graphics (TOG), vol. 39, no. 4, pp. 75–1, 2020.

[107] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma, "Multi-memory convolutional neural network for video super-resolution," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2530–2544, 2019.

[108] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 3, pp. 933–948, 2019.

[109] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "Mucan: Multi-correspondence aggregation network for video super-resolution," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. Springer, 2020, pp. 335–351.

[110] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[111] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3360–3369.

[112] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," 2018.

[113] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. Mu Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[114] X. Yang, W. Xiang, H. Zeng, and L. Zhang, "Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4781–4790.

[115] R. Wang, X. Liu, Z. Zhang, X. Wu, C.-M. Feng, L. Zhang, and W. Zuo, "Benchmark dataset and effective inter-frame alignment for real-world video super-resolution," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1168–1177.

[116] J. Xiao, H. Yong, and L. Zhang, "Degradation model learning for real-world single image super-resolution," in Proceedings of the Asian Conference on Computer Vision, 2020.

[117] K. Zhang, J. Liang, L. V. Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," 2021.