Spatial-Temporal Space Hand-in-Hand: Spatial-Temporal Video Super-Resolution via Cycle-Projected Mutual Learning

Mengshun Hu[1,2†] Kui Jiang[1,2†] Liang Liao[3] Jing Xiao[1,2] Junjun Jiang[4] Zheng Wang[1,2‡]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University [2]Hubei Key Laboratory of Multimedia and Network Communication Engineering
[3]Nanyang Technological University [4]Harbin Institute of Technology

Abstract

Spatial-Temporal Video Super-Resolution (ST-VSR) aims to generate super-resolved videos with higher resolution (HR) and higher frame rate (HFR). Quite intuitively, pioneering two-stage based methods complete ST-VSR by directly combining two sub-tasks: Spatial Video Super-Resolution (S-VSR) and Temporal Video Super-Resolution (T-VSR) but ignore the reciprocal relations among them. Specifically, 1) T-VSR to S-VSR: temporal correlations help accurate spatial detail representation with more clues; 2) S-VSR to T-VSR: abundant spatial information contributes to the refinement of temporal prediction. To this end, we propose a one-stage based Cycle-projected Mutual learning network (CycMu-Net) for ST-VSR, which makes full use of spatial-temporal correlations via the mutual learning between S-VSR and T-VSR. Specifically, we propose to exploit the mutual information among them via iterative up-and-down projections, where the spatial and temporal features are fully fused and distilled, helping the high-quality video reconstruction. Besides extensive experiments on benchmark datasets, we also compare our proposed CycMu-Net with S-VSR and T-VSR tasks, demonstrating that our method significantly outperforms state-of-the-art methods. Codes are publicly available at: https://github.com/hhhhhumengshun/CycMuNet.

1. Introduction

Spatial-temporal video super-resolution (ST-VSR) aims to produce the high-resolution (HR) and high-frame-rate (HFR) video sequences from the given low-resolution (LR) and low-frame-rate (LFR) input. This task has drawn great attention due to its popular applications [29, 30, 53], including HR slow-motion generation, movie production, high-definition television upgrades, *etc*. Great success has been recently achieved in ST-VSR tasks, as illustrated in Figure 1(a), which can be roughly divided into two categories: two-stage and one-stage based methods. The former decomposes it into two sequential sub-tasks: spatial video super-resolution (S-VSR) and temporal video super-resolution (T-VSR), which are individually completed with image/video super-resolution technologies [19, 51, 58] and video frame interpolation technologies [28, 40]. However, more spatial information generated by the S-VSR task can be used for the refinement of temporal prediction, while more temporal information predicted by the T-VSR task can be used to facilitate the reconstruction of spatial details. As a result, the two-stage based approaches are far from producing satisfied predictions due to lacking the ability to mutually explore the coupled correlations between S-VSR and T-VSR.

Recently, integrating these two sub-tasks into a unified framework with a one-stage process becomes more popular. Naturally, based on the parallel or serially processing modes (Figure 1(b) (i) for parallel process and (ii)(iii) for serial process), diverse and effective schemes have been developed [7, 8, 29, 30, 53, 55]. Unfortunately, the parallel methods [29, 30] barely consider the coupled correlations between the two sub-tasks, while the serial methods [53,55] fail to fully exploit mutual relations since they only focus on the unilateral relationship, such as "T-to-S" or "S-to-T". In particular, the unilateral learning will accumulate reconstruction errors, which we define as cross-space (spatial and temporal spaces) errors, consequently leading to obvious aliasing effect in super-resolved results.

For thorough utilization of spatial and temporal information, we propose to promote the one-stage method with mutual learning, and devise a novel cycle-projected mutual learning network (CycMu-Net) for ST-VSR. As shown in Figure 1(c), the philosophy of CycMu-Net is to explore the mutual relations and achieve the spatial-temporal fusion to eliminate the cross-space errors. Specifically, the key part of CycMu-Net is the iterative up-and-down projection units between the spatial and temporal embedding spaces, involv-

[†]Equal Contribution
[‡]Corresponding Author

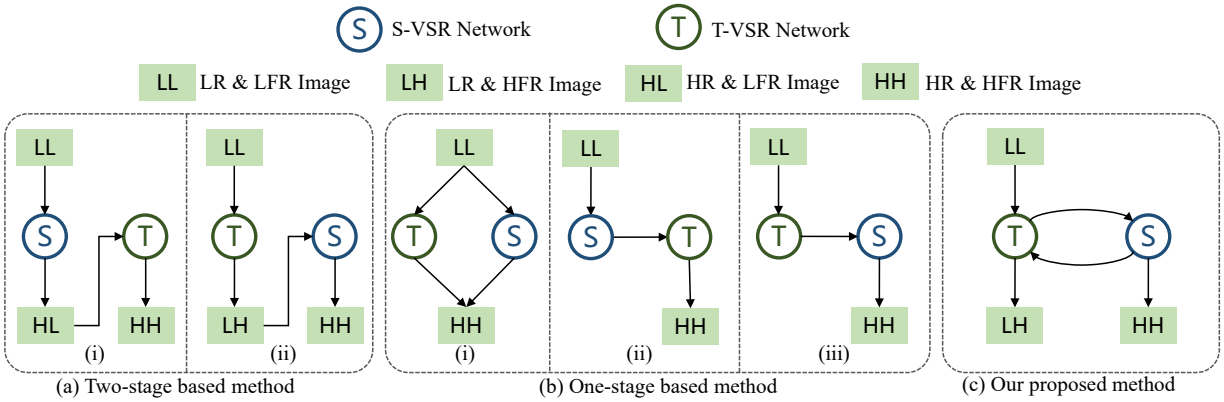

Figure 1. Different schemes for ST-VSR. (a) Two-stage based methods: (i) they perform ST-VSR task by independently using the advanced S-VSR methods and then T-VSR methods or vice versa (ii). (b) One-stage based method: they unify S-VSR and T-VSR tasks into one model with parallel or cascaded manners without considering the mutual relations between S-VSR and T-VSR. (c) Mutual method: **Our method makes full use of the mutual relations via mutual learning between S-VSR and T-VSR.**

ing a process of aggregating temporal relations to achieve an accurate representation of spatial details, and a feedback refinement of temporal information via the updated spatial prediction. We validate the proposed CycMu-Net on the ST-VSR task and its two sub-tasks, involving S-VSR and T-VSR. Experimental results demonstrate that CycMu-Net achieves appealing improvements over the SOTA methods on all tasks. Our contributions are summarized as follows:

1) We propose a novel one-stage based cycle-projected mutual learning network (CycMu-Net) for spatial-temporal video super-resolution, which can make full use of the coupled spatial-temporal correlations via mutual learning between S-VSR and T-VSR.

2) To eliminate the cross-space errors and promote the inference accuracy, we devise iterative up-and-down projection units to exploit the mutual information between S-VSR and T-VSR for a better spatial-temporal fusion. In these units, more spatial information are provided for the refinement of temporal prediction while temporal correlations are used to promote texture and detail reconstruction.

3) We conduct extensive experiments on ST-VSR, S-VSR and T-VSR tasks for a comprehensive evaluation, showing that our method performs well against the state-of-the-art methods.

2. Related Work

2.1. Spatial Video Super-Resolution

S-VSR aims to super-resolve LR frames to HR frames with temporal alignment and spatial fusion. Thus, the key to this task lies in fully exploiting temporal correlations among multiple frames. Some methods perform temporal alignment using explicit motion estimation (*e.g.*, optical flow) and then fuse all aligned reference frames for S-VSR [3, 6, 42, 47, 50, 56]. However, optical flow estimation is error-prone, which may degrade the S-VSR performance [34]. To address this issue, some methods propose to apply deformable convolution to sample more spatial pixels based on multiple motion offsets [13, 61] for implicit alignment [7, 49, 51]. It is effective but time-consuming, since the alignment is required for all reference frames each time when super-resolving the target frame. Other researchers propose to explore the global temporal correlations with recurrent networks that propagate inter-frame information forward and backward independently [8, 26, 53, 55]. However, extra motion estimation networks are still required to assist the recurrent network based S-VSR approach in dealing with large and complex motions [53, 55].

2.2. Temporal Video Super-Resolution

T-VSR (*i.e.*, video frame interpolation) aims to generate the non-existent intermediate frame between two consecutive frames. The key to this task is to find correspondences between consecutive frames to synthesize intermediate frames. The popular T-VSR methods mainly fall into two categories: kernel-based and flow-based methods. The former implicitly aligns the input frames by learning the dynamic convolution kernels, which are used to resample the input frames to produce intermediate frames [11, 18, 33, 39, 40, 44]. Due to only resampling the local neighborhood patches, the aforementioned methods usually lead to ambiguous results. By contrast, the latter first estimates bidirectional optical flows between two consecutive frames and then warps to synthesize the intermediate frames based on the predicted optical flows [2, 3, 24, 25, 28, 37, 38]. While achieving impressive progress, they rely heavily on the accuracy of current advanced optical flow algorithms [27, 41, 46, 48].

2.3. Spatial-Temporal Video Super-Resolution

ST-VSR technologies tend to increase spatial and temporal resolution of LR and LFR videos [22, 30, 53, 55].

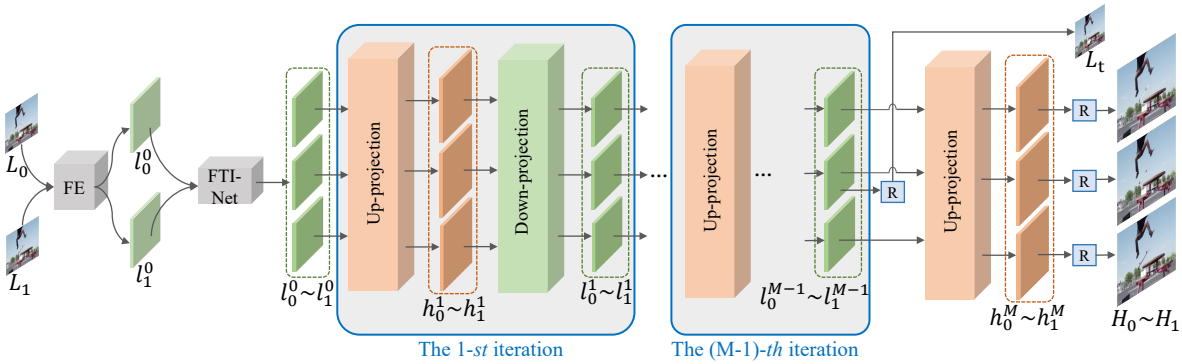

Figure 2. **Architecture of the proposed Cycle-projected Mutual learning network (CycMu-Net).** Given two LR input frames, we first extract representations from input frames by feature extractor (FE) and obtain an initialized intermediate representation by feature temporal interpolation network (FTI-Net). We then adopt mutual learning to exploit the mutual information between S-VSR and T-VSR and obtain M $2\times$ HR and LR representations via M up-projection units and $M-1$ down-projection units. Finally, we concatenate and feed the multiple $2\times$ HR representations and LR representations into reconstruction network (R) to reconstruct corresponding HR images and LR intermediate frame, respectively.

For example, Shechtman *et al.* adopt a directional spatial-temporal smoothness regularization to constrain high spatial-temporal resolution video reconstruction [43]. Mudenagudi *et al.* [36] formulate their ST-VSR method as a posteriori-Markov Random Field [17] and optimize it by achieving the Maximum of graph-cuts [5]. However, the above methods cost great computational consumption and fail to model complex spatial-temporal correlations. Recently, learning-based methods attempt to unify S-VSR and T-VSR into a single-stage framework for ST-VSR. Kim *et al.* utilize a multi-scale U-net to learn ST-VSR based on a multi-scale spatial-temporal loss [30]. Haris *et al.* propose to explore spatial-temporal correlations by a pre-trained optical flow model for frame interpolation and refinement [22]. Xiang *et al* devise a unified framework to interpolate intermediate features by deformable convolution [51], explored global temporal correlations by bidirectional deformable ConvLSTM [54], and finally reconstructed high spatial-temporal videos by a reconstruction network [53]. Inspired by [53], Xu *et al.* introduce a locally temporal feature comparison module to extract local motion cues in videos, achieving better performance on various datasets [55]. However, as shown in Figure 1(b), the mutual relations between S-VSR and T-VSR are under-explored, while leading to the accumulated reconstruction errors. To address this issue, we propose a cycle-projected mutual learning network that learns the spatial-temporal correlations via the iterative operation of spatial and temporal fusion (S-VSR and T-VSR) during the forward propagation and backward optimization.

2.4. Mutual Learning

Mutual learning is to make a pool of untrained students to learn collaboratively and teach each other for solving the task [59]. Dual-NMT utilizes mutual learning to teach two cross-lingual translation models each other interactively machine translation [23]. Tanmay Batra *et al.* propose to learn multiple models jointly and communicate object attributes each other for recognising the same set of object categories [4]. Dong *et al.* adopted this tool to exploit non-adjacent features for image dehazing by fusing features from different levels [15]. The closest thing to our work is DBPN [19], which proposes utilize mutually iterative up- and down-sampling layers to learn nonlinear relationships between LR and HR images to guide the image SR task. Previous studies have validated the effectiveness of mutual learning techniques for low-level tasks [14,16,21,60]. However, the existing methods tend to exploit the mutual learning to refine the mapping relations of different scale spaces ("LR-to-HR" and "HR-to-LR"). Inspired by them, we introduce a novel cycle-projected mutual learning mechanism to cooperatively characterise the spatial and temporal feature representations.

3. Cycle-Projected Mutual Learning Network

In this section, we first provide an overview of the proposed Cycle-projected Mutual learning network (CycMu-Net) for ST-VSR. As shown in Figure 2, given two LR input frames L_0 and L_1, our goal is to synthesize HR intermediate frame H_t and the corresponding HR input frames H_0 and H_1 ($2\times$, $4\times$, or $8\times$). In addition, we also generate a LR frame L_t as a intermediate result. The proposed CycMu-Net first extracts the representation from the input frames by a feature extractor (FE). To synthesize the initialized LR intermediate representation, we introduce a cascading multi-scale architecture as our feature temporal interpolation network (FTI-Net), designed to learn bi-directional motion offsets to handle complex motions and interpolate intermediate representation by deformable convolution. To make full use of the mutual relations ("T-to-S" or "S-to-T")

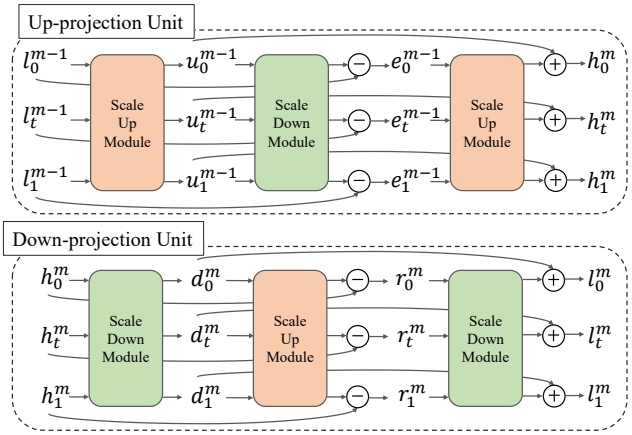

Figure 3. Illustration of the proposed **up-projection unit (UPU)** and **down-projection unit (DPU)** in the CycMu-Net.

between S-VSR and T-VSR, we adopt mutual learning that temporal correlations contribute to accurate spatial representations and updated spatial predictions refine temporal information via feedback, to eliminate the cross-space errors, which can be achieved via iterative up-projection units (UPUs) and down-projection units (DPUs). After several iterations, we obtain multiple HR and LR representations and then concatenate them into the reconstruction network (R) to generate the corresponding HR images H_0, H_t and H_1 ($2\times$, $4\times$, or $8\times$) and LR image L_t.

3.1. Cycle-Projected Mutual Learning

Inspired by [19] that adequately addressed the mutual dependencies of low- and high-resolution images via mutually connected up- and down-sampling layers, in this paper, we propose a new mutual learning model including iterative UPUs and DPUs to explore the mutual relations between S-VSR and T-VSR. In particular, temporal correlations provide more clues to compensate detailed spatial representation via UPUs while abundant spatial details are used to refine the temporal predictions via DPUs.

As shown in the top of Figure 3, the UPU captures temporal correlations for S-VSR. We firstly project previous LR temporal representations l_0^{m-1}, l_t^{m-1} and l_1^{m-1} to corresponding HR representations u_0^{m-1}, u_t^{m-1} and u_1^{m-1} based on a scale up module, which can be described as follows:

$$[u_0^{m-1}, u_t^{m-1}, u_1^{m-1}] = UP_0([l_0^{m-1}, l_t^{m-1}, l_1^{m-1}]), \quad (1)$$

where $UP_0(\cdot)$ denotes the scale up module. It first performs multi-frame progressive fusion by fusion resblocks [57], which implicitly exploit intra-frame spatial correlations and inter-frame temporal correlations, then upsamples each feature by bilinear interpolation and 1×1 convolution. $m = 1, 2..., M$ denotes the number of UPU.

Then we try to project the super-resolved representations back to LR representations and compute the corresponding residuals (errors) e_0^{m-1}, e_t^{m-1} and e_1^{m-1} between back-projected representations and original LR representations, respectively, which can be defined as follows:

$$\begin{aligned}[e_0^{m-1}, e_t^{m-1}, e_1^{m-1}] = DN([u_0^{m-1}, u_t^{m-1}, u_1^{m-1}]) \\ - [l_0^{m-1}, l_t^{m-1}, l_1^{m-1}],\end{aligned} \quad (2)$$

where $DN(\cdot)$ denotes the scale down module. It first reduces the input to the original input resolution via 4×4 convolution with stride 2, and then further implicitly explores intra-frame spatial correlations and inter-frame temporal correlations of LR representations by fusion resblocks [57].

Finally, we project residual representations again back to HR representations (back-project) and eliminate the corresponding original super-resolved representations errors (cross-space errors) to obtain the final super-resolution outputs of the unit by

$$\begin{aligned}[h_0^m, h_t^m, h_1^m] = UP_1([e_0^{m-1}, e_t^{m-1}, e_1^{m-1}]) \\ + [u_0^{m-1}, u_t^{m-1}, u_1^{m-1}],\end{aligned} \quad (3)$$

where $UP_1(\cdot)$ denotes the scale up module.

As shown in the bottom of Figure 3, the procedure for DPU is very similar, while its main role is to obtain refined LR temporal representations by projecting the previously updated HR representations, which can provide abundant spatial details. (Please refer to the supplementary materials for more details about formula proof, scale up module and scale down module)

3.2. Spatial-Temporal Video Super-Resolution

The overall framework of CycMu-Net is shown in Figure 2, consisting of the following sub-modules: feature extraction network, feature temporal interpolation network, multiple up-projection units, multiple down-projection units, and reconstruction network. Specifically, we extract representations among multiple frames via feature extraction network (FE) and interpolate the intermediate representations via the feature temporal interpolation network (FTI-Net). Then we use the proposed multiple UPUs and DPUs to obtain multiple LR and HR representations with the mutual learning. Finally, the reconstruction network (R) generates LR intermediate frame and HR intermediate frames by concatenating all LR and HR representations. Below we describe the details of each sub-module.

Feature temporal interpolation network. Deformable convolution [13, 61] has been shown to be effective for video frame interpolation [10] and video super-resolution [49]. Some methods extended deformable convolution and explored a wider range of offsets by employing a multi-scale framework to handle feature alignment for small and large displacements [51, 53, 55]. Inspired by them, we utilize a cascading multi-scale architecture for our feature

temporal interpolation network (FTI-Net) to estimate the bi-directional motion offsets from input frames. Along with the motion offsets estimation, we adopt deformable convolution to interpolate forward and backward representations from the missing intermediate frames. To blend these two representations for obtaining an initial intermediate representation, we use the two learnable convolution kernels to estimate the weights, which can adaptively fuse the two representations according to their importance. (More details on FTI-Net are provided in the supplementary materials)

Reconstruction network. After the mutual relations between S-VSR and T-VSR are exploited by the proposed iterative up-and-down projections, we concatenate and feed multiple HR representations into convolution layers to reconstruct the corresponding HR frames. In addition, we also reconstruct a LR intermediate frame based on multiple LR representations. To optimize the whole CycMu-Net, we use a reconstruction loss function:

$$\begin{aligned}\mathcal{L}_r = \lambda_1\rho(L_t - L_t^{GT}) + \lambda_2\rho(H_t - H_t^{GT}) \\ + \lambda_3\rho(H_0 - H_0^{GT}) + \lambda_4\rho(H_1 - H_1^{GT}),\end{aligned} \tag{4}$$

where L_t^{GT}, H_0^{GT}, H_t^{GT} and H_1^{GT} refer to the corresponding ground-truth video frames. $\rho(x) = \sqrt{x^2 + \omega^2}$ is the Charbonnier penalty function [9, 32]. We set the constant ω and weights λ_1, λ_2, λ_3 and λ_4 to 10^{-3}, 1, 1, 0.5 and 0.5, respectively.

3.3. Implementation Details

We implement the proposed CycMu-Net using Pytorch 1.9 with four NVIDIA 2080Ti and optimize the model using AdaMax optimizer [31] with a momentum of 0.9. The batch size is set to 10 with image resolution of 64×64. The initial learning rate is set to 4×10^{-4} and reduced by a factor of 10 every 20 epochs for a total of 70 epochs. We compare HR intermediate frame H_t for the evaluation of ST-VSR. In addition, we also compare our proposed CycMu-Net with S-VSR and T-VSR methods, where 4× HR frame H_0 and LR intermediate frame L_t are used for the evaluations of S-VSR and T-VSR, respectively.

4. Experimental Results

4.1. Datasets and Metrics

Vimeo90k [56]. We use Vimeo90K dataset to train our proposed CycMu-Net. This dataset consists of many triplets with different scenes from 14,777 video clips with image resolution of 448×256. Among them, 51,312 triplets and 3,782 triplets are used for training and testing, respectively. In order to increase the diversity of data, we use horizontal and vertical flipping or reverse the order of input frames for data augmentation. For a fair comparison with other algorithms during training, we downscale to original images to 64×64 with Bicubic interpolation for 2× and 4× SR, and downscaled to original images to 32×32 with Bicubic interpolation for 8× SR.

UCF101 [45]. The UCF101 dataset consists of videos with a large variety of human actions. There are 379 triplets with the resolution of 256×256 for testing. The original images are sampled to 32×32, 64×64 and 128×128 with Bicubic for 8×, 4× and 2× SR tasks in testing.

Middlebury [1]. The Middlebury dataset is widely used to evaluate video frame interpolation algorithms [2, 10]. Here, we select Other set which provides the ground-truth middle frames, only to test our method on T-VSR task. The image resolution in this dataset is around 640×480 pixels.

Metric. We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [52] and the average Interpolation Error (IE) for performance evaluation. The higher PSNR and SSIM and lower IE values indicate better super-resolution and interpolation performance.

4.2. Comparisons with State-of-the-Art Methods

ST-VSR. We compare our CycMu-Net with state-of-the-art two-stage and one-stage based ST-VSR methods. For the two-stage based ST-VSR methods, SepConv [40], AdaCoF [33] and CAIN [12] are introduced for T-VSR task, while Bicubic Interpolation, RBPN [20], DBPN [19] and EDVR [51] are used for S-VSR. For one-stage based ST-VSR methods, we compare our CycMu-Net with Zooming SlowMo [53], STARnet [22] and TMNet [55]. For fair comparison, three triplets from **Vimeo90K** dataset are used to retrain SlowMo and TMNet methods.

Quantitative results. Quantitative results are presented in Table 1. We can see that besides fewer parameters, one-stage based methods show significant superiority than the two-stage based methods in all metrics. In particular, the best two-stage based method (SepConv+RBPN) is 0.66dB lower than our method for 8× VSR on Vimeo90K dataset. Furthermore, compared to the state-of-the-art one-stage based methods, our proposed CycMu-Net outperforms STARNet [22], Zooming Slow-Mo [53] and TMNet [55] on all datasets with all metrics, while with only one-tenth of parameters to STARnet. All these results validate the effectiveness of our proposed method for ST-VSR task.

Qualitative results. The qualitative results of seven ST-VSR baselines with their PSNR and SSIM values are shown in Figure 4. As expected, two-stage based ST-VSR methods tend to produce blurry results (see the yellow boxes) since they ignore the mutual relations between S-VSR and T-VSR, which help the accurate texture inference. Compared to two-stage based methods, one-stage based ST-VSR methods can generate complete results. However, these methods ignore that S-VSR provides abundant spatial information for the refinement of temporal prediction, leading to the generated results without more texture information

T-VSR Method	S-VSR Method	UCF101			Vimeo90K			UCF101			Vimeo90K			UCF101			Vimeo90K			Parameters (millions)
		PSNR	SSIM	IE	PSNR	SSIM	IE	PSNR	SSIM	IE	PSNR	SSIM	IE	PSNR	SSIM	IE	PSNR	SSIM	IE	
SepConv [40]	Bicubic	29.988	0.944	4.531	30.628	0.937	4.234	26.189	0.874	7.154	27.287	0.866	6.582	22.877	0.779	11.201	24.181	0.782	9.989	21.7
SepConv [40]	DBPN [19]	32.041	0.958	3.729	32.179	0.955	3.415	28.380	0.915	5.573	28.969	0.903	5.268	25.135	0.845	8.298	26.016	0.834	7.717	21.7+10.4
SepConv [40]	RBPN [20]	31.859	0.957	3.795	32.377	0.958	3.300	28.650	0.920	5.400	29.507	0.914	4.912	25.323	0.823	8.067	26.409	0.846	7.275	21.7+12.7
SepConv [40]	EDVR [51]	—	—	—	—	—	—	28.650	0.920	5.388	29.481	0.914	4.909	—	—	—	—	—	—	21.7+20.7
AdaCoF [33]	Bicubic	30.056	0.945	4.458	30.760	0.936	4.203	26.187	0.874	7.133	27.243	0.864	6.624	22.877	0.778	11.193	24.160	0.781	10.029	21.8
AdaCoF [33]	DBPN [19]	32.167	0.958	3.630	32.341	0.954	3.401	28.557	0.917	5.430	29.214	0.903	5.207	25.164	0.845	8.253	25.935	0.832	7.804	21.8+10.4
AdaCoF [33]	RBPN [20]	31.997	0.958	3.692	32.537	0.957	3.288	28.840	0.922	5.237	29.584	0.914	4.865	25.349	0.851	8.026	26.155	0.841	7.466	21.8+12.7
AdaCoF [33]	EDVR [51]	—	—	—	—	—	—	28.848	0.923	5.226	29.700	0.916	4.810	—	—	—	—	—	—	21.8+20.7
CAIN [12]	Bicubic	29.931	0.941	4.627	30.578	0.931	4.412	25.987	0.865	7.456	26.908	0.851	7.035	22.505	0.743	12.166	23.820	0.759	10.691	42.8
CAIN [12]	DBPN [19]	31.741	0.954	3.904	31.796	0.946	3.819	27.814	0.901	6.105	28.100	0.877	6.125	23.672	0.779	10.561	24.764	0.784	9.478	42.8+10.4
CAIN [12]	RBPN [20]	31.721	0.955	3.896	31.980	0.949	3.702	27.995	0.906	5.930	28.377	0.887	5.855	23.566	0.781	10.498	24.605	0.787	9.437	42.8+12.7
CAIN [12]	EDVR [51]	—	—	—	—	—	—	28.339	0.911	5.711	28.690	0.893	5.642	—	—	—	—	—	—	42.8+20.7
STARnet [22]		—	—	—	—	—	—	28.829	0.920	—	30.608	0.926	—	—	—	—	—	—	—	111.6
Zooming Slow-Mo [53]		32.200	0.959	3.630	33.270	0.963	2.982	28.931	0.923	5.184	30.621	0.927	4.354	25.376	0.850	8.054	26.829	0.851	7.018	11.1
TMNet [55]		32.211	0.960	3.620	33.298	0.964	2.974	28.988	0.924	5.149	30.699	0.929	4.311	25.424	0.852	7.984	26.994	0.854	6.874	12.3
CycMu-Net		32.258	0.960	3.608	33.545	0.965	2.885	29.020	0.925	5.130	30.750	0.929	4.287	25.486	0.853	7.931	27.062	0.856	6.827	11.1

Table 1. Quantitative comparisons ($\times 2$, $\times 4$, $\times 8$ from left to right) of the state-of-the art methods for ST-VSR. The numbers in red and blue represent the best and second best performance.

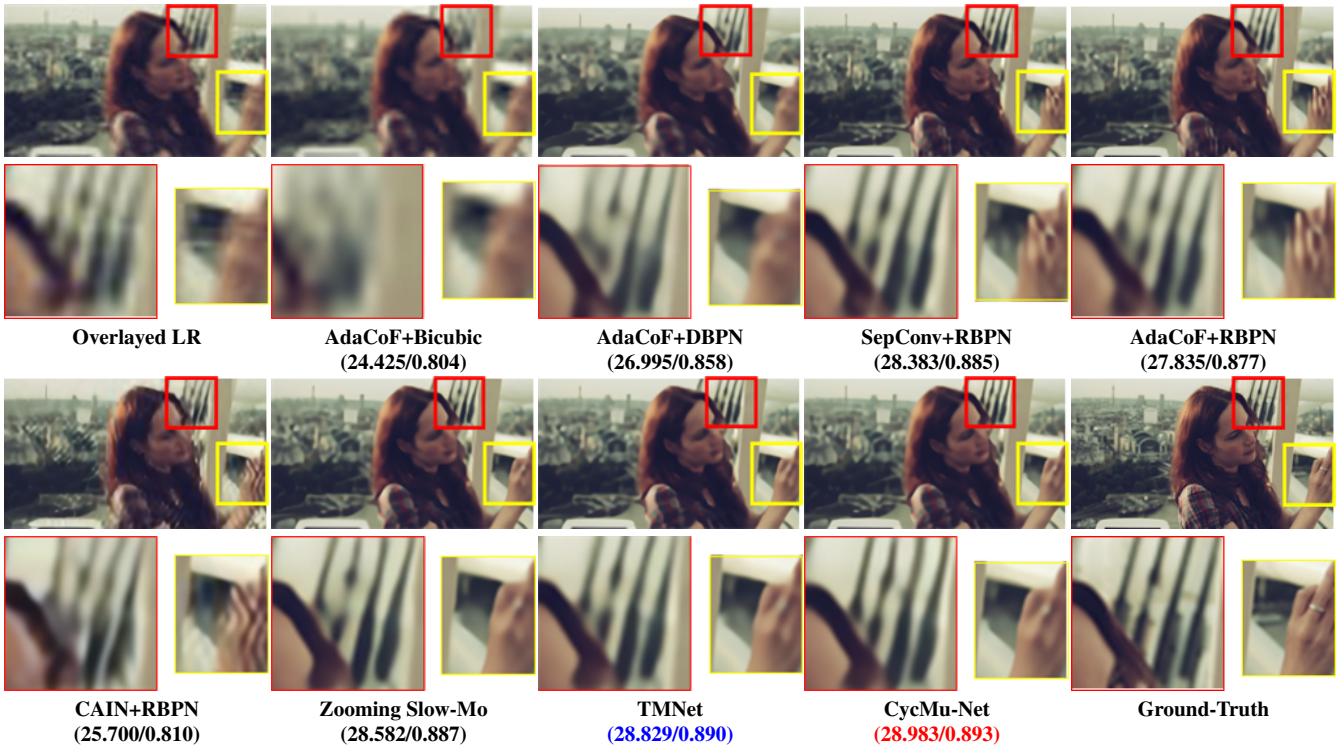

Figure 4. Visual comparisons ($8\times$) with state-of-the-art methods on **Vimeo90K** dataset.

Methods	UCF101			Vimeo90K			Parameters (millions)
	PSNR	SSIM	IE	PSNR	SSIM	IE	
Bicubic	27.254	0.889	6.232	28.135	0.879	5.994	—
DBPN [19]	30.898	0.938	4.211	31.484	0.928	4.137	10.4
RBPN [20]	31.309	0.943	4.035	32.417	0.939	3.759	12.7
EDVR [51]	31.452	0.944	3.974	32.558	0.941	3.680	20.7
CycMu-Net	31.463	0.944	3.980	32.472	0.940	3.735	11.1

Table 2. Quantitative comparisons of the state-of-the art methods for S-VSR (H_0) on UCF101 and Vimeo90K datasets.

(see red and yellow boxes). On the contrary, our proposed method explores the mutual relations between S-VSR and T-VSR, which contribute to generating sharper results with clearer structure and texture. (More visual comparisons are provided in the supplementary materials)

S-VSR. We compare the proposed network with image SR methods including Bicubic and DBPN [19], and S-VSR methods including RBPN [20] and EDVR [51]. The results on S-VSR are shown in Table 2, showing that S-VSR methods (EDVR [51] and RBPN [20]) can achieve superior performance than image SR methods (bicubic and DBPN [19]) by referring to multiple frames for temporal correlations. In addition, we can see that our CycMu-Net has comparable results with EDVR, but it requires only half of the parameters of EDVR and three triplets rather than seven frames for training. This also validates the powerful generalization ability of our network, and our proposed up-projection units are helpful for S-VSR tasks by exploiting temporal correlations from T-VSR.

T-VSR. We compare our proposed network with state-of-the-art T-VSR which include SpeConv-L_f [40], SepConv-

Overlayed LR | **EDSC** (1.826) | **DAIN** (1.868) | **AdaCoF++** **(1.812)** | **CAIN** (3.054) | **CycMu-Net** **(1.539)** | **Ground-truth**

Figure 5. Visual comparisons of temporal video super-resolution on **Middlebury** dataset.

Overlayed LR | **Model (a)** **(27.182/0.887)** | **Model (b)** **(27.276/0.893)** | **Model (c)** **(27.528/0.897)** | **Model (d)** **(27.673/0.900)** | **Ground-truth**

Figure 6. Visual comparisons (4×) of four variants for the ablation studies on **Vimeo90K** dataset.

Methods	UCF101			Vimeo90K			MB-Other	Parameters
	PSNR	SSIM	IE	PSNR	SSIM	IE	IE	(millions)
SpeConv-L_f [40]	37.883	0.982	2.264	36.506	0.985	1.936	1.355	21.6
SpeConv-L_1 [40]	37.953	0.983	2.221	36.788	0.986	1.845	1.310	21.6
EDSC [11]	37.946	0.983	2.271	37.326	0.988	1.824	1.302	8.9
DAIN [2]	38.172	0.983	2.131	36.686	0.986	1.862	1.346	24.0
CyclicGen++ [35]	37.644	0.981	2.261	33.935	0.973	2.660	1.750	19.8
AdaCoF++ [33]	38.387	0.983	2.088	36.874	0.987	1.857	1.304	21.8
CAIN [12]	35.407	0.979	2.849	34.857	0.979	2.729	2.369	42.8
CycMu-Net	38.850	0.984	2.012	39.074	0.990	1.422	0.983	11.1

Table 3. Quantitative comparisons of the state-of-the art methods for T-VSR (L_t).

L_1 [40], EDSC [11], DAIN [2], CyclicGen++ [35], AdaCoF++ [33] and CAIN [12]. The results on T-VSR are shown in Table 3. We can find that our proposed method is significantly better than the state-of-the-art video frame interpolation. For example, PSNR values of our proposed CycMu-Net are 1.1dB and 1.6dB higher than EDSC [11] on UCF101 and Vimeo90K datasets, respectively. In addition, we show the visualized results and IE value from four temporal video super-resolution method in Figure 5, our proposed method produces intermediate frame with more details (*e.g.*, the shoe). We attribute this to the fact that when we train the ST-VSR network, we make full use of HR information from S-VSR via down-projection units. Therefore, the interpolated frame can obtain more texture and detailed information from S-VSR.

4.3. Model Analysis

Ablation Study. To further verify the key modules in CycMu-Net, comprehensive ablation studies are conducted for 4× SR.

Methods	FTI		PU		UCF101			Vimeo90K		
	FFI	DFI	PP	CP	PSNR	SSIM	IE	PSNR	SSIM	IE
Model (a)	✓				28.861	0.922	5.243	30.170	0.921	4.616
Model (b)		✓			28.926	0.924	5.161	30.510	0.926	4.415
Model (c)		✓	✓		28.940	0.924	5.150	30.544	0.926	4.390
Model (d)		✓		✓	28.996	0.924	5.144	30.650	0.928	4.338

Table 4. Quantitative comparisons on the performance (4×) of different modules. FTI denotes feature temporal interpolation, FFI denotes fusion feature interpolation, DFI denotes deformable feature interpolation, PU denotes projection units. PP denotes plain-projected units and CP denotes cycle-projected units.

Model (a): A fusion feature interpolation (FFT) network is used to direct fuse input information from input frames and produce intermediate representation without motion estimation. Then two pixel-shuffle layers take the representations as inputs, and produce the 4× SR video with a convolution.
Model (b): We add deformable convolution as implicit motion estimation into feature interpolation network (FTI-Net) in Model (a) as our deformable feature interpolation (DFI) network, as stated in section 3.2
Model (c): Based on Model (b), we add addition iterative plain-projection units (PP) without up-down sampling between the feature temporal interpolation network and reconstruction network.
Model (d): The complete version of CycMu-Net.

The visual and numerical comparisons are shown in Figure 6 and Table 4. Compared to Model (a) that produces the intermediate representations without motion estimation, the results of Model (b) show that adopting deformable convo-

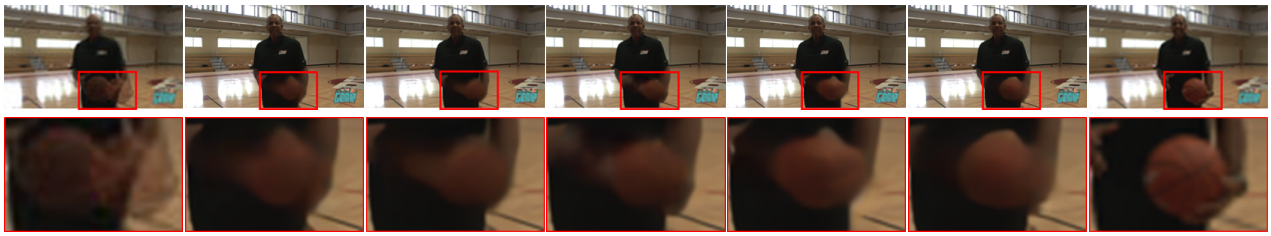

Figure 7. Visual comparisons (4×) of different numbers of up-projection and down-projection units for the ablation studies on **Vimeo90K** dataset.

M	UCF101			Vimeo90K			Parameters
	PSNR	SSIM	IE	PSNR	SSIM	IE	(millions)
2	28.939	0.923	5.181	30.480	0.926	4.420	7.3
4	28.982	0.924	5.149	30.601	0.927	4.360	9.2
6	29.020	0.925	5.130	30.750	0.929	4.287	11.1
8	29.030	0.925	5.130	30.753	0.929	4.282	13.0
10	29.044	0.925	5.128	30.791	0.929	4.273	14.9

Table 5. Quantitative comparisons on the performance (4×) of different number of projection units.

lution for implicit frame interpolation can bring 0.34dB gain on Vimeo90K dataset and improves the visual result (*e.g.*, the edge of the moving car). Based on Model (b), the addition of plain projection units (Model (c)) can help Model (b) to generate a car with clearer structure. Unfortunately, they fail to recover key details (*e.g.*, license plate). On the contrary, our proposed Model (d) can generate more credible SR results. It demonstrates the fact that our proposed up- and down-projection units eliminate cross-space errors while plain-projection units magnify errors.

Impacts of Up-projection and Down-projection Units. To demonstrate the effectiveness of our up-projection units and down-projection units, we construct multiple networks ($M = 2, 4, 6, 8, 10$) by setting different numbers of projection units. The visual and numerical results on 4 × are shown in Figure 7 and Table 5. As the numbers of up-projection and down-projection units increase, CycMu-Net produces results with more complete structure and details (*e.g.*, the basketball), and achieves better results in term of PSNR, SSIM and IE on two datasets. Considering the trade-off between efficacy and efficiency, we set M to 6 to predict the final results of the proposed CycMu-Net. These also verify that the proposed up-projection and down-projection units play important roles in mutually benefiting from S-VSR and T-VSR. In addition, in order to analyze the specific role of the projection units that temporal correlations are exploited to promote the texture and detail information. In Figure 8, it is shown that each up-projection unit generates feature map, which contains different types of HR components and increases the quality of S-VSR. This demonstrates that multiple up-projection units can obtain diverse HR representations for guiding the better super-resolution reconstruction.

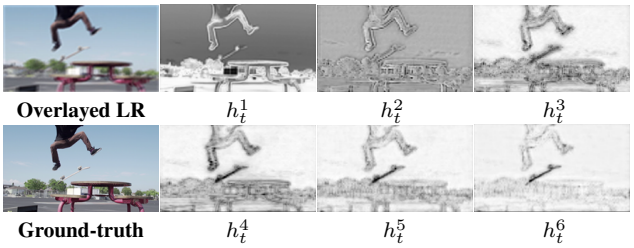

Figure 8. Feature maps from up-projection units in CycMu-Net where $M = 6$. Each feature map has been visualized using same grayscale colormap.

5. Conclusion

In this work, we propose a novel one-stage based Cycle-projected Mutual learning network (CycMu-Net) for spatial-temporal video super-resolution. Theoretically, we introduce mutual learning to explore the interactions between spatial video super-resolution (S-VSR) and temporal video super-resolution (T-VSR), from which the abundant spatial information and temporal correlations are aggregated to infer accurate intermediate frame. Specifically, an elaborate iterative representation between up-projection units and down-projection units is introduced to make full use of the spatial-temporal features while eliminating the inference errors. Extensive experiments demonstrate our proposed method performs well against the state-of-the-art methods in both S-VSR, T-VSR and ST-VSR tasks. While achieving impressive performance, one limitation of this study is that since videos might contain dramatically changing scenes, the spatial-temporal correlations of large motion or SR factors is hardly predicted via the iterative up-projection and down-projection units. One reasonable scheme is to alleviate the learning burden by dividing it into multiple sub-tasks with small motion, which is helpful for accurate texture inference.

Acknowledgements. This work was supported by National Key R&D Project (2021YFC3320301) and National Natural Science Foundation of China (62171325). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011. 5

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, pages 3703–3712, 2019. 2, 5, 7

[3] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE TPAMI*, 2019. 2

[4] Tanmay Batra and Devi Parikh. Cooperative learning with visual attributes. *arXiv preprint arXiv:1705.05512*, 2017. 3

[5] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE TPAMI*, 23(11):1222–1239, 2001. 3

[6] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, pages 4778–4787, 2017. 2

[7] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4:3, 2020. 1, 2

[8] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, pages 4947–4956, 2021. 1, 2

[9] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, volume 2, pages 168–172. IEEE, 1994. 5

[10] Xianhang Cheng and Zhenzhong Chen. Video frame interpolation via deformable separable convolution. In *AAAI*, volume 34, pages 10607–10614, 2020. 4, 5

[11] Xianhang Cheng and Zhenzhong Chen. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE TPAMI*, 2021. 2, 7

[12] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *AAAI*, volume 34, pages 10663–10671, 2020. 5, 6, 7

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2, 4

[14] Shengyang Dai, Mei Han, Ying Wu, and Yihong Gong. Bilateral back-projection for single image super resolution. In *ICME*, pages 1039–1042. IEEE, 2007. 3

[15] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020. 3

[16] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Nonlocal back-projection for adaptive image enlargement. In *ICIP*, pages 349–352. IEEE, 2009. 3

[17] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE TPAMI*, (6):721–741, 1984. 3

[18] Shurui Gui, Chaoyue Wang, Qihua Chen, and Dacheng Tao. Featureflow: Robust video interpolation via structure-to-texture generation. In *CVPR*, pages 14004–14013, 2020. 2

[19] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for single image super-resolution. *CVPR*, 2019. 1, 3, 4, 5, 6

[20] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, pages 3897–3906, 2019. 5, 6

[21] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projectinetworks for single image super-resolution. *IEEE TPAMI*, 43(12):4323–4337, 2020. 3

[22] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *CVPR*, pages 2859–2868, 2020. 2, 3, 5, 6

[23] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *NIPS*, 29, 2016. 3

[24] Mengshun Hu, Liang Liao, Jing Xiao, Lin Gu, and Shin'ichi Satoh. Motion feedback design for video frame interpolation. In *ICASSP*, pages 4347–4351. IEEE, 2020. 2

[25] Mengshun Hu, Jing Xiao, Liang Liao, Zheng Wang, Chia-Wen Lin, Mi Wang, and Shin'ichi Satoh. Capturing small, fast-moving objects: Frame interpolation via recurrent motion enhancement. *IEEE TCSVT*, 2021. 2

[26] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE TPAMI*, 40(4):1015–1028, 2017. 2

[27] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, pages 8981–8989, 2018. 2

[28] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, pages 9000–9008, 2018. 1, 2

[29] Jaeyeon Kang, Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Deep space-time video upsampling networks. In *ECCV*, pages 701–717. Springer, 2020. 1

[30] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Fisr: deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *AAAI*, volume 34, pages 11278–11286, 2020. 1, 2, 3

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, pages 624–632, 2017. 5

[33] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *CVPR*, pages 5316–5325, 2020. 2, 5, 6, 7

[34] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, pages 335–351. Springer, 2020. 2

[35] Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *AAAI*, volume 33, pages 8794–8802, 2019. 7

[36] Uma Mudenagudi, Subhashis Banerjee, and Prem Kumar Kalra. Space-time super-resolution using graph-cut optimization. *IEEE TPAMI*, 33(5):995–1008, 2010. 3

[37] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *CVPR*, pages 1701–1710, 2018. 2

[38] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, pages 5437–5446, 2020. 2

[39] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *CVPR*, pages 670–679, 2017. 2

[40] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, pages 261–270, 2017. 1, 2, 5, 6, 7

[41] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, pages 4161–4170, 2017. 2

[42] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, pages 6626–6634, 2018. 2

[43] Eli Shechtman, Yaron Caspi, and Michal Irani. Increasing space-time resolution in video. In *ECCV*, pages 753–768. Springer, 2002. 3

[44] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. Video frame interpolation via generalized deformable convolution. *IEEE TMM*, 2021. 2

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[46] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 2

[47] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *CVPR*, pages 4472–4480, 2017. 2

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 2

[49] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, pages 3360–3369, 2020. 2, 4

[50] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through hr optical flow estimation. In *ACCV*, pages 514–529. Springer, 2018. 2

[51] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, pages 0–0, 2019. 1, 2, 3, 4, 5, 6

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 5

[53] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *CVPR*, pages 3370–3379, 2020. 1, 2, 3, 4, 5, 6

[54] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015. 3

[55] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *CVPR*, pages 6388–6397, 2021. 1, 2, 3, 4, 5, 6

[56] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127(8):1106–1125, 2019. 2, 5

[57] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, pages 3106–3115, 2019. 4

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 1

[59] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 3

[60] Yang Zhao, Rong-Gang Wang, Wei Jia, Wen-Min Wang, and Wen Gao. Iterative projection reconstruction for fast and efficient image upsampling. *Neurocomputing*, 226:200–211, 2017. 3

[61] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 2, 4