Spatio-Temporal Transformer Network for Video Restoration

Tae Hyun Kim[1,2], Mehdi S. M. Sajjadi[1,3], Michael Hirsch[1,4†], Bernhard Schölkopf[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
{tkim,msajjadi,bs}@tue.mpg.de
[2] Hanyang University, Seoul, Republic of Korea
[3] Max Planck ETH Center for Learning Systems
[4] Amazon Research, Tübingen, Germany
hirsch@amazon.com

Abstract. State-of-the-art video restoration methods integrate optical flow estimation networks to utilize temporal information. However, these networks typically consider only a pair of consecutive frames and hence are not capable of capturing long-range temporal dependencies and fall short of establishing correspondences across several timesteps. To alleviate these problems, we propose a novel Spatio-temporal Transformer Network (STTN) which handles multiple frames at once and thereby manages to mitigate the common nuisance of occlusions in optical flow estimation. Our proposed STTN comprises a module that estimates optical flow in both space and time and a resampling layer that selectively warps target frames using the estimated flow. In our experiments, we demonstrate the efficiency of the proposed network and show state-of-the-art restoration results in video super-resolution and video deblurring.

Keywords: Spatio-temporal transformer network, Spatio-temporal flow, Spatio-temporal sampler, Video super-resolution, Video deblurring

1 Introduction

Motion estimation via dense optical flow is a crucial component in video processing including restoration tasks such as video super-resolution and video deblurring. While traditional approaches try to register consecutive frames within a video sequence through energy minimization [1–3], more recent works [4–8] have demonstrated that convolutional neural networks (CNNs) enable accurate, fast and reliable optical flow estimation by end-to-end deep learning.

The success and applicability of CNNs in optical flow estimation has been made possible through the availability of large amounts of data. State-of-the-art approaches [4–6] are based on supervised learning and require labeled data with ground truth optical flow. However, as such networks are trained on synthetic

[†]The scientific idea and a preliminary version of the code were developed at the MPI prior to joining Amazon.

datasets with known ground truth [9, 10], their generalization to real-world data remains challenging due to the intrinsic differences between training and test data and the limited variability of synthetic imagery [8].

To circumvent the need for labeled training data with ground truth optical flow, recent work [7, 11, 12] has promoted unsupervised learning by introducing resampling layers that are differentiable and allow end-to-end training. Such a resampling layer would warp an image to a reference frame according to the estimated optical flow, such that the measured pixel-wise distance in image rather than optical flow space can be used as a training objective.

In this paper, we build on these ideas and propose a task-specific end-to-end unsupervised approach for *Spatio-temporal Flow* estimation which is dense optical flow that selectively captures long-range temporal dependencies by allowing for several consecutive frames as network input. To this end, we extend the Spatial Transformer Network [13] to a *Spatio-temporal Transformer Network* (STTN) which is able to establish dense pixel correspondences across space and time. We show that reasoning over several consecutive frames and choosing one of them per pixel location helps mitigate the commonly known problem of occlusions in optical flow estimation. A further advantage of our approach is that it can be trained in an unsupervised fashion and thus renders the availability of large labeled data unnecessary. When used in conjunction with a video restoration network tailored for a specific task, we obtain a substantial performance gain with minimal computational overhead. We demonstrate the effectiveness of our proposed STTN for the challenging tasks of video super-resolution and video deblurring, and improve upon the state-of-the-art by a substantial margin. In summary, we make the following contributions:

- We introduce a spatio-temporal flow estimation network which selectively captures long-range temporal dependencies without a large computational overhead and which alleviates the occlusion problem in conventional optical flow estimation.
- We present a spatio-temporal sampler which enables spatio-temporal manipulation of the input data by using the estimated spatio-temporal flow.
- We show promising results on challenging video restoration tasks such as video super-resolution and deblurring by simply placing the proposed network on top of the state-of-the-art methods.

2 Related Work

2.1 Optical Flow Estimation

Many computer vision tasks such as tracking, 3D reconstruction, and video restoration rely on accurate optical flow and as a result, flow estimation techniques have been widely studied. Traditional optical flow estimation methods are based on energy optimization, and minimize the proposed energy models with various optical flow constraints. These generative models focus on studying better prior models [14–17] and more accurate likelihood models [18–20],

and these approaches perform well on the Middlebury optical flow dataset [21] where motion displacements are relatively small. With the release of more challenging datasets beyond the experimental Middlebury dataset, such as the MPI Sintel [22] and KITTI [23] datasets, more robust algorithms to handle large displacements have been studied [24–27].

As it has recently become possible to generate large synthetic flow datasets [4], deep learning based flow estimation approaches are being actively studied. In particular, Dosovitskiy *et al.* [4] propose FlowNet, the first neural approach which directly renders optical flow from two consecutive images. As FlowNet is fully convolutional, it can be trained in an end-to-end manner and exhibits real-time performance. In follow-up studies, Ilg *et al.* [5] extend FlowNet and improve flow accuracy by passing flow through a stacked architecture which includes multiple sub-networks specialized for both small and large displacements, and Ranjan *et al.* [6] propose a faster network which can handle large displacements by embedding the traditional coarse-to-fine approach into the flow estimation network.

However, as it is difficult to collect large amounts of labeled flow datasets for real scenes, there are some works that train the flow estimation networks in an unsupervised manner. Ren *et al.* [7] introduce an unsupervised learning method for flow networks which minimizes the traditional energy function with a data term based on photometric consistency coupled with a smoothness term. Meister *et al.* [8] improve the flow accuracy by training the network with robust census transform which is more reliable at occluded regions. As there is a considerable difference between synthetic flow datasets and the real-world videos used in our restoration tasks, these unsupervised learning methods which train flow networks on datasets of real scenes are more closely related to our work than supervised learning methods.

2.2 Video Restoration

Video Super-Resolution. Since spatial alignment is a key element for high-quality video super-resolution, many conventional methods seek to find correspondences among adjacent frames by optimizing energy models for the enhancement task [1, 3, 28]. Moreover, learning-based state-of-the-art video super-resolution methods are also composed of an alignment mechanism and a super-resolution network. After aligning several previous and future frames to the current frame, all frames are fed into a super-resolution network which then combines the information from several views. This method is often applied in a sliding window over the video [29–33]. While older methods use classical approaches for the alignment [29], recent approaches employ neural networks for this task as well. While Caballero *et al.* [30] warp frames using a dense optical flow estimation network before feeding the result to a super-resolution network, Makansi *et al.* [31] combine warping and mapping to high-resolution space into a single step. Although using a larger number of previous and future frames leads to higher-quality results, the computational overhead of aligning several frames rises linearly with the number of inputs, limiting the amount of information that can be combined into a single output frame. A recent approach by

Sajjadi *et al.* [34] therefore uses a frame-recurrent architecture that reuses the previous output frame for the next frame which leads to higher efficiency and quality compared to sliding-window approaches. However, for higher efficiency, only a single previous frame is fed into the network for the next iteration, leading to suboptimal results in case of occlusions.

Video Deblurring. Early approaches to remove blur in video frames are based on image deblurring techniques which remove uniform blurs caused by translational camera shake. Cai *et al.* [35] and Zhang *et al.* [36] study sparsity characteristics of the latent sharp frames and the blur kernels to restore the uniformly blurred images. For the next step, to remove non-uniform blurs caused by rotational camera shake, several methods tackle the simultaneous registration and restoration problems [37–40]. In contrast, Wulff and Black[41] propose a method which jointly segments and restores the differently blurred foreground and background regions by object and ego motions, and Kim *et al.* [2, 42] propose methods which remove spatially varying blurs without relying on accurate segmentation results by parameterizing the blur kernel using optical flow [43].

As large motion blur datasets become available [42, 44–46], several deep learning approaches have been proposed to restore the video frames. First, Shuochen *et al.* [45] propose a deep neural network taking a stack of neighboring blurry frames as input. As these frames are aligned with the reference frame, the proposed network can easily exploit multiple frames and reconstruct sharp images. Recently, Wieschollek *et al.* [46] introduce a recurrent neural network which can handle an arbitrary number of temporal inputs as well as arbitrary spatial input sizes. Unlike all of the previous deblurring methods which require significant time to restore a video frame, Kim *et al.* [47] propose a fast online video deblurring method by efficiently increasing the receptive field of the network without adding a computational overhead to handle large motion blurs. Moreover, by training the network to enforce temporal smoothness, their method achieves state-of-the-art results with near real-time performance.

3 Proposed Method

Spatial transformer networks (STN) proposed by Jaderberg *et al.* [13] that enable generic warping of the feature maps are widely used in numerous vision applications. In particular for video restoration tasks, many deep learning approaches are based on variants of STN to estimate optical flow between adjacent frames and to align the target frames onto the reference frame [30–32, 34]. However, the STN only allows spatial manipulation of the input data. To handle multiple video frames at each time step, one needs to employ STN multiple times which is a severe limitation when applied in real-time settings.

We therefore introduce a novel spatio-temporal transformer network (STTN) which efficiently enables spatio-temporal warping of the input data and alleviates the limitations of the conventional STN without a large computational overhead. In Fig. 1, the overall structure of our proposed STTN which is composed of a

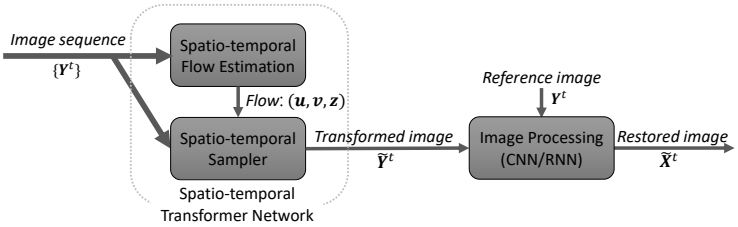


Fig. 1. Our spatio-temporal transformer network is composed of a flow estimation network which calculates spatio-temporal flow and a sampler which selectively transforms the multiple target frames to the reference. An image processing network follows for the video restoration task.

spatio-temporal flow estimation network and a spatio-temporal sampler is illustrated. In the following sections, we describe the details of each component.

3.1 Spatio-Temporal Flow Estimation Network

Traditional model-based approaches solve optical flow estimation and video restoration problems jointly [1–3, 28], and recent learning-based approaches estimate the optical flow by using off-the-shelf methods [29, 45] or by employing sub-networks to estimate flow [30, 31, 34].

However, all of these previous methods estimate optical flow between two consecutive frames (reference and target frames), and thus it is required to calculate flow N times to handle N target frames for each time step. Moreover, as shown in Fig. 3, conventional flow estimation networks are unreliable where correspondences are not well established (e.g., occlusion and illumination change).

To overcome these limitations, we propose a new spatio-temporal flow estimation network which takes a sequence of multiple neighboring frames $\{\mathbf{Y}^t\} \in \mathbb{R}^{H\times W\times C\times T}$ as input where H, W, C, and T denote the height, width, number of channels and number of input frames, and outputs a normalized three-dimensional spatio-temporal flow $(\mathbf{u}, \mathbf{v}, \mathbf{z}) \in [-1, 1]^{H\times W\times 3}$. Notably, the height and width of the output flow can be different from those of input depending on applications. Therefore, our spatio-temporal network can handle multiple frames very efficiently at a single time step, and it becomes more robust to occlusions and illumination changes since there are *multiple* matching candidates from multiple target frames, unlike conventional works which consider only one target frame.

The detailed configuration of the proposed U-net [48] like spatio-temporal flow estimation network is shown in Fig. 2. All convolutional layers are performed with 3x3 filters and are followed by batch normalization [49] and ReLu except for the last convolutional layer which is followed by *tanh* to output a normalized flow. As our flow estimation network is fully convolutional, it can be used to handle frames of arbitrary (spatial) size at inference time once trained.

Layer	Filter size	Stride	Output shape
C0	$3 \times 3 \times (C \times T) \times 64$	1	$H/1 \times W/1 \times 64$
C1	$3 \times 3 \times 64 \times 64$	1	$H/1 \times W/1 \times 64$
C2	$3 \times 3 \times 64 \times 64$	1	$H/1 \times W/1 \times 64$
C3	$3 \times 3 \times 64 \times 128$	2	$H/2 \times W/2 \times 128$
C4	$3 \times 3 \times 128 \times 128$	1	$H/2 \times W/2 \times 128$
C5	$3 \times 3 \times 128 \times 256$	2	$H/4 \times W/4 \times 256$
C6	$3 \times 3 \times 256 \times 256$	1	$H/4 \times W/4 \times 256$
Upscaling	nearest neighbor, 2x	-	$H/2 \times W/2 \times 256$
C7+C4	$3 \times 3 \times 256 \times 128$	1	$H/2 \times W/2 \times 128$
C8	$3 \times 3 \times 128 \times 128$	1	$H/2 \times W/2 \times 128$
Upscaling	nearest neighbor, 2x	-	$H/1 \times W/1 \times 128$
C9+C2	$3 \times 3 \times 128 \times 64$	1	$H/1 \times W/1 \times 64$
C10	$3 \times 3 \times 64 \times 64$	1	$H/1 \times W/1 \times 64$
C11	$3 \times 3 \times 64 \times 3$	1	$H/1 \times W/1 \times 3$

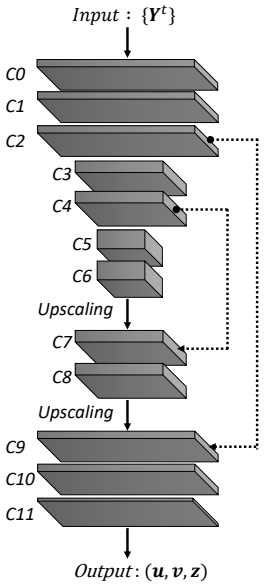


Fig. 2. Our fully convolutional spatio-temporal flow estimation network takes multiple frames as input ($H \times W \times C \times T$) and renders a spatio-temporal flow ($H \times W \times 3$) as output.

3.2 Differentiable Spatio-Temporal Sampler

To synthesize a new image aligned with the reference frame by selectively warping the multiple target frames using the spatio-temporal flow in Sec. 3.1, a new sampler which performs sampling in three-dimensional spatio-temporal space is required. In this paper, we propose a spatio-temporal sampler by naturally extending the conventional spatial sampling module from two-dimensional to three-dimensional space. Our spatio-temporal sampler interpolates intensity values from multiple target frames as:

$$\tilde{\mathbf{Y}}^{t}_{(x,y)} = \sum_{n}^{W} \sum_{m}^{H} \sum_{i \in \Delta} \mathbf{Y}^{t+i}_{(n,m)} \cdot \delta(\mathbf{u}_{(x,y)}, \mathbf{v}_{(x,y)}, \mathbf{z}_{(x,y)}), \tag{1}$$

where $\tilde{\mathbf{Y}}_{(x,y)}$ denotes the interpolated pixel value at location (x, y) and $\mathbf{Y}^{t+i}_{(n,m)}$ is the intensity value of $\mathbf{Y}^{t+i}$ at a pixel location (n, m) with temporal shift $i \in \Delta$. For example, we can define a sliding window of the form $\Delta = \{-2, \ldots, 3\}$. The function δ defines an interpolation method using the spatio-temporal flow $(\mathbf{u}, \mathbf{v}, \mathbf{z})$. Any function δ whose sub-gradient is defined could be used for sampling as introduced in [13]. Here, we employ trilinear interpolation for δ in our video

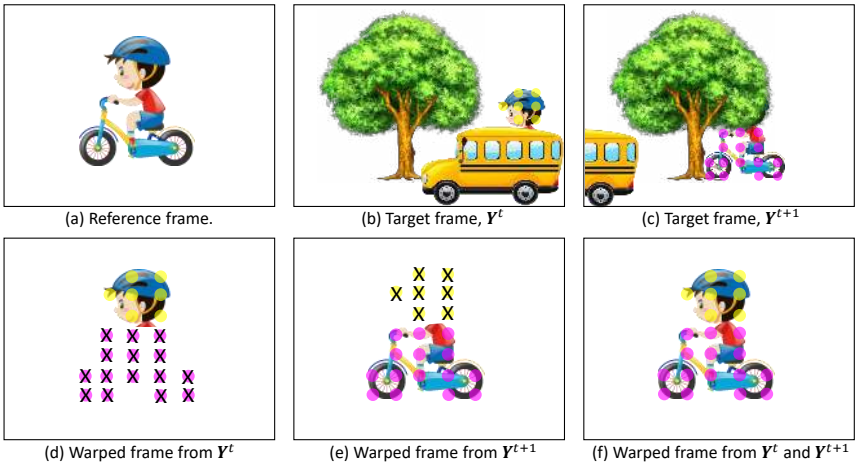


Fig. 3. Spatial transformer network vs. spatio-temporal transformer network. Conventional spatial transformer networks require to be employed multiple times to extract information separately as in (d) and (e), and they are unreliable in areas of occlusions. In contrast, the proposed spatio-temporal transformer can exploit multiple frames at the same time and find a single best match among multiple target frames.

restoration tasks. It is given by

$$\begin{aligned}\tilde{\mathbf{Y}}^t_{(x,y)} = &\sum_n^W \sum_m^H \sum_{i\in\Delta} \mathbf{Y}^{t+i}_{(n,m)} \cdot \max(0, 1-|x+\mathbf{u}_{(x,y)}-n|) \\ &\cdot \max(0, 1-|y+\mathbf{v}_{(x,y)}-m|) \cdot \max(0, 1-|\mathbf{z}_{(x,y)}-(t+i)|).\end{aligned} \tag{2}$$

Note that the unnormalized version of spatio-temporal flow $(\mathbf{u}, \mathbf{v}, \mathbf{z})$ is used in (2), that is, $\mathbf{u}_{(x,y)}$ and $\mathbf{v}_{(x,y)}$ denote the horizontal and vertical motion displacements, and $\mathbf{z}_{(x,y)}$ is mapped to a real value close to a target frame index which is favored to be matched at (x, y).

Similarly to the bilinear spatial sampling procedure in [13], our trilinear sampling mechanism in three-dimensional space is also differentiable. The gradient with respect to our spatio-temporal flow is derived as follows:

$$\begin{aligned}\frac{\partial \tilde{\mathbf{Y}}^t_{(x,y)}}{\partial \mathbf{z}_{(x,y)}} = &\sum_n^W \sum_m^H \sum_{i\in\Delta} \mathbf{Y}^{t+i}_{(n,m)} \cdot \max(0, 1-|x+\mathbf{u}_{(x,y)}-n|) \\ &\cdot \max(0, 1-|y+\mathbf{v}_{(x,y)}-m|) \cdot \begin{cases} 0, & \text{if } |\mathbf{z}_{(x,y)}-(t+i)| \geq 1 \\ 1, & \text{if } \mathbf{z}_{(x,y)} \leq t+i \\ -1, & \text{otherwise} \end{cases}\end{aligned} \tag{3}$$

Note that the gradients for $\frac{\partial \tilde{\mathbf{Y}}^t_{(x,y)}}{\partial \mathbf{u}_{(x,y)}}$ and $\frac{\partial \tilde{\mathbf{Y}}^t_{(x,y)}}{\partial \mathbf{v}_{(x,y)}}$ can be derived similarly. More generally, our spatio-temporal transformer can take a set of feature maps $\boldsymbol{U}^t$ as

input rather than images $\mathbf{Y}^t$. The gradient with respect to $\boldsymbol{U}^t$ is then given as:

$$\begin{aligned}\frac{\partial \tilde{\mathbf{Y}}^t_{(x,y)}}{\partial \mathbf{U}^t_{(n,m)}} = &\sum_n^W \sum_m^H \sum_{i \in \Delta} \max(0, 1 - |x + \mathbf{u}_{(x,y)} - n|) \\ &\cdot \max(0, 1 - |y + \mathbf{v}_{(x,y)} - m|) \cdot \max(0, 1 - |\mathbf{z}_{(x,y)} - (t+i)|).\end{aligned} \tag{4}$$

This means that the proposed spatio-temporal sampler allows the loss gradients to backpropagate readily into the input frames or feature maps.

3.3 Spatio-Temporal Transformer Network

Our spatio-tempral flow estimation network which handles multiple different frames at the same time can replace multiple usages of optical flow estimation modules in conventional approaches [30, 31, 45] with much less computational effort. Moreover, our spatio-temporal sampling mechanism can also be also processed very efficiently with modern GPUs. Additionally, as the proposed network can find a better corresponding point from several frames in spatio-temporal space rather than the conventional STN approach which estimates the matching point in a single target frame. This leads to a network which is much more robust against outliers as a result of occlusion or illumination changes.

Our spatio-temporal transformer network directly extends the spatial transformer network [13] into three-dimensional space. Because of this, many characteristics of the previous spatial transformer network can be generalized in the proposed network. First, our spatio-temporal transformer network can be easily trained in an end-to-end manner as loss gradients can flow backwards through both the sampler and the flow estimation network, and it can be placed in any location of the conventional networks to selectively transform or merge multiple feature maps efficiently. As a result, the proposed module can be used in numerous applications apart from our video restoration tasks. Second, unlike the spatial transformer which enables to upscale and downscale the feature maps only in the two-dimensional spatial domain, our spatio-temporal transformer allows to change shapes not only in the spatial domain but also in temporal space. Next, as suggested in [13], our network can also be added multiple times at increasing depths of a network or in parallel to handle multiple objects at different time steps while the spatial transformer network can handle multiple objects only at a single time step.

Unsupervised Spatio-Temporal Flow Learning. Recent learning-based optical flow estimation methods are trained on large synthetic datasets such as Flyinging Chairs [4] and the MPI Sintel dataset [22]. However, to the best of our knowledge, there is no available dataset which can be used to train our spatio-temporal flow estimation network directly, and it is not straightforward to utilize optical flow datasets to train the proposed network. Therefore, we train our network in an unsupervised manner. Particularly, for our video restoration

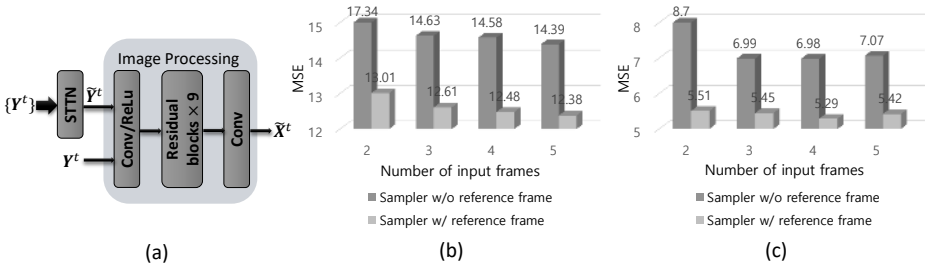


Fig. 4. Image reconstruction performance is evaluated on super-resolution and deblurring datasets for a varying number of frames. (a) Video restoration network with STTN. (b) Reconstruction loss on video super-resolution dataset. (c) Reconstruction loss on video deblurring dataset.

applications, we propose a loss to train our flow estimation network by constraining the synthesized image from our spatio-temporal sampler as:

$$L_{flow} = \|\tilde{\mathbf{Y}}^t - \mathbf{X}^t\|^2, \tag{5}$$

where $\mathbf{X}^t$ denotes the ground truth frame corresponding to the reference frame at time step t.

4 Experiments

In this section, we demonstrate the power and versatility of our STTN and show how state-of-the-art image and video restoration networks can be further improved with the simple addition of the proposed spatio-temporal transformer.

4.1 Ablation Study

To evaluate the warping performance of the spatio-temporal transformer network, we trained the approach using video datasets with various settings of the hyperparameters (e.g., the number of target frames). As shown in Fig. 4 (a), we use a video restoration network with STTN, and the image processing module which is composed of convolutional layers and residual blocks as used in [34, 44, 50]. The network is trained by jointly minimizing L_{flow} in (5) and MSE between the latent and ground truth images, and we compare the warped (synthesized) and ground truth frames.

First, we train the network using a super-resolution dataset. Since no standard high-resolution video super-resolution dataset is available, we have collected a set of high-quality youtube videos and extracted 120k UHD frames to train the networks. As a next step, we have generated low-resolution video frames by downscaling the clean video frames by a factor of 4 and subsequently quantizing them before we upscale the low-resolution frames to the original image size

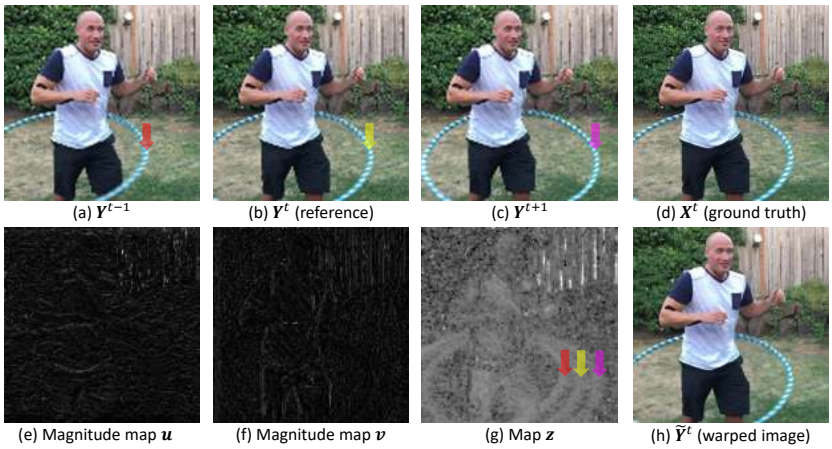


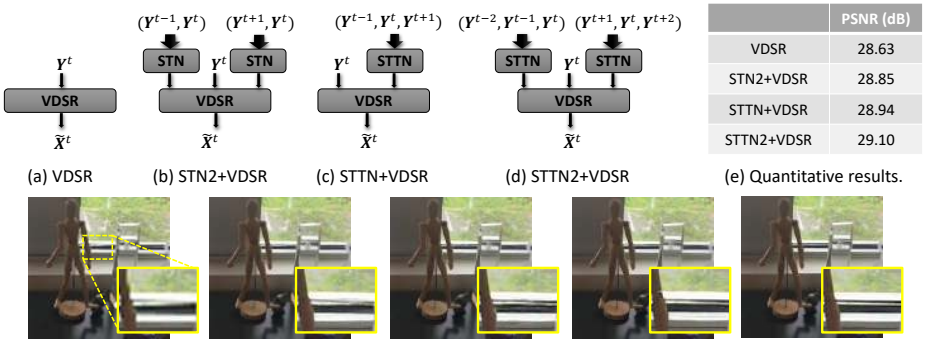
Fig. 5. (a)-(c) Input frames. (d) Ground truth frame. (e-f) Magnitude of motion displacements. (g) $\mathbf{Y}^{t-1}$ is favored whenever $0 \leq \mathbf{z} < 0.33$, and $\mathbf{Y}^t$ is favored whenever $0.33 \leq \mathbf{z} < 0.66$. Otherwise, $\mathbf{Y}^{t+1}$ is chosen. (h) Transformed image by our sampler.

again. We evaluate eight networks trained with different settings: Four of them take two to five input frames and then perform sampling with the estimated flow and target frames. The other four networks also take two to five frames as network input, but run the sampler with the reference frame as well as target frames, that is, the reference frame is also considered as a target frame. Similarly, we also compare different networks trained on video deblurring datasets [45]. In Fig. 4 (b)–(c), the quality of the warped frame $\tilde{\mathbf{Y}}^t$ is evaluated in terms of reconstruction error (i.e., L_{flow}). Overall, the networks tend to give better results with more inputs, though the gain in performance slowly saturates as we add more frames. Moreover, we observe that the reconstruction error is significantly reduced by considering the reference frame as a target, since it can then render the reference frame itself where correspondences are not available. Therefore, we consider the reference frame as a target in our network for subsequent experiments. In Fig. 5 (e)–(h), our flow maps and the transformed images are visualized. As expected, the occluded background regions by the moving hula hoop ring, indicated by colored arrows, are mainly mapped by the reference frame itself.

4.2 Video Super-resolution

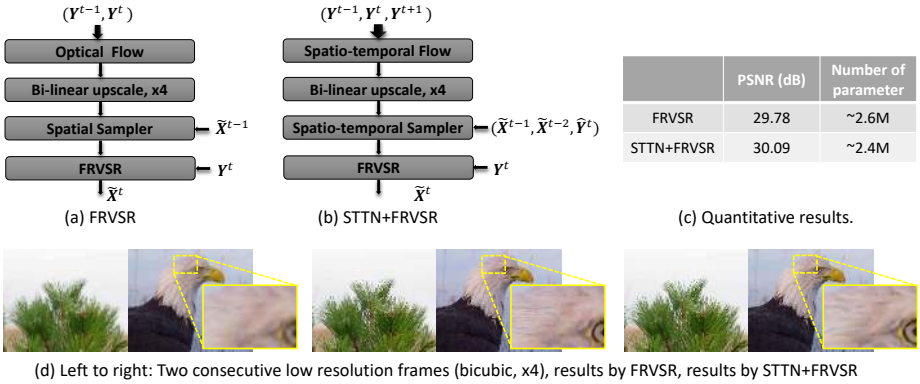
We further integrate our network into state-of-the-art super-resolution networks and provide comparisons to demonstrate the performance of our network.

Comparison with VDSR [51]. Kim *et al.* proposed VDSR in Fig. 6 (a) which raised the bar for single image super-resolution [51]. We show how VDSR can



	PSNR (dB)
VDSR	28.63
STN2+VDSR	28.85
STTN+VDSR	28.94
STTN2+VDSR	29.10

Fig. 6. Comparisons with VDSR [51].



	PSNR (dB)	Number of parameter
FRVSR	29.78	~2.6M
STTN+FRVSR	30.09	~2.4M

Fig. 7. Comparisons with FRVSR [34]. In (b), $\hat{\mathbf{Y}}^t$ denotes the bilinearly upscaled (x4) version of the reference frame $\mathbf{Y}^t$.

be naturally improved and extended to handle multiple video frames using our spatio-temporal transformer network.

In our experiments, STN2+VDSR in Fig. 6 (b) refers to the conventional model which integrates two spatial transformer networks to calculate bi-directional optical flow as in [30, 31]. Our STTN+VDSR in Fig. 6 (c) has one spatio-temporal transformer, and similarly to STN2+VDSR, two spatio-temporal transformer networks are placed in our STTN2+VDSR model in Fig. 6 (d). Note that for fair comparisons, we modify our spatio-temporal flow estimation network in Fig. 2 to calculate optical flow (output shape: $H \times W \times 3 \rightarrow H \times W \times 2$), and use the flow in STN2+VDSR. All these networks are trained on the video super-resolution dataset used in Sec. 4.1 using the ADAM optimizer (learning rate 10^{-4}). We used 256x256 patches with a minibatch size of 8, and the networks are trained for 300k iterations. Moreover, STN2+VDSR, STTN+VDSR

and STTN2+VDSR were trained by minimizing the loss

$$Loss = \|\tilde{\mathbf{X}}^t - \mathbf{X}^t\|^2 + L_{flow}, \tag{6}$$

and quantitative comparisons are given in Fig. 6 (e). Notably, STTN+VDSR and STTN2+VDSR outperform VDSR, and STTN+VDSR shows competitive results compared to STN2+VDSR while using half as much computation in flow estimation. Qualitative results in Fig. 6 (f) show that the proposed models restore thin edges better, consistent with the quantitative results.

Comparison with FRVSR [34]. To improve the performance of previous video super-resolution networks which do not utilize previously computed results, Sajjadi *et al.* [34] propose FRVSR which is based on a recurrent architecture, outperforming previous state-of-the-art methods. However, FRVSR only utilizes a single previously restored frame as shown in Fig. 7 (a) as utilizing several frames would be computationally expensive. To handle multiple frames efficiently, we integrate our spatio-temporal transformer into FRVSR as shown in Fig. 7 (b). We refer to the resulting model as STTN+FRVSR.

Using the dataset used in Sec. 4.1, FRVSR and STTN+FRVSR are trained for 500k iterations following the training procedure suggested by [34]. To train our STTN+FRVSR, we use the loss

$$Loss = \sum_t \|\tilde{\mathbf{X}}^t - \mathbf{X}^t\|^2 + L_{flow}. \tag{7}$$

In Fig. 7 (c), the networks are quantitatively evaluated in terms of PSNR, and the proposed network achieves better results compared to FRVSR with fewer parameters. Moreover, in Fig. 7 (d), visual comparisons are provided when there is a shot change between two consecutive frames. While FRVSR produces artifacts by propagating wrong details from the previous frame after the shot change, our STTN+FRVSR renders better result since STTN+FRVSR can use the reference frame itself for synthesizing $\tilde{\mathbf{Y}}^t$.

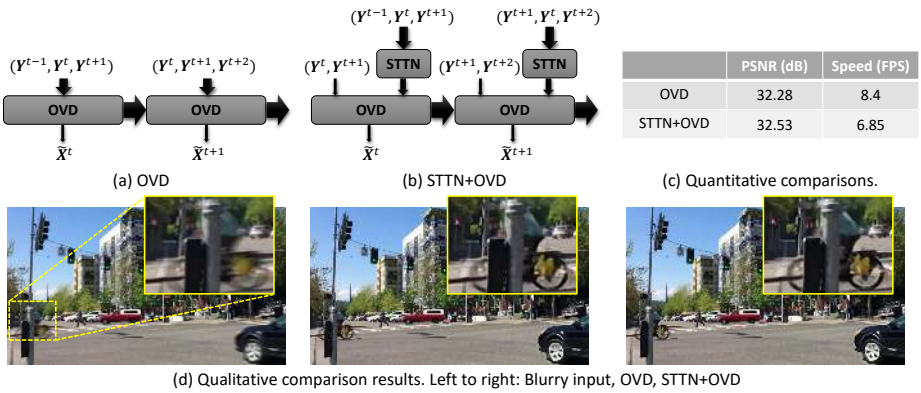
4.3 Video Deblurring

To further show the versatility of STTN, we also embed our network into stat-of-the-art video deblurring networks and provide comparison results.

Comparison with DVD [45]. The deep video deblurring (DVD) network proposed by Shuochen *et al.* [45] removes spatially varying motion blurs in video frames. More specifically, as shown in Fig. 8 (a), DVD takes a stack of five adjacent frames as network input (one reference and four target frames) and then aligns the target frames with the reference frame using an off-the-shelf optical flow method [52]. It thus requires optical flow calculations four times at each time step, slowing down the method. To reduce the computation time in flow calculation and to further improve the deblurring performance, we propose two different models on top of DVD.

	PSNR (dB)	Speed (FPS)
DVD	31.02	0.04
STTN+DVD	31.45	1.99
RSTTN+DVD	31.61	1.99

Fig. 8. Comparisons with DVD [45]. A function W in (a) warps the target frame to the reference frame.



	PSNR (dB)	Speed (FPS)
OVD	32.28	8.4
STTN+OVD	32.53	6.85

Fig. 9. Comparisons with OVD [47].

First, our STTN+DVD model in Fig. 8 (b) calculates spatio-temporal flow and synthesizes a selectively transformed image from five consecutive frames. Next, the DVD network takes both the synthesized image and the reference frame as input. Our RSTTN+DVD model in Fig. 8 (c) also takes five frames at each time step, but two of them are previously restored sharp frames ($\tilde{X}^{t-2}$, $\tilde{X}^{t-1}$). RSTTN+DVD is thus a kind of recurrent network, but does not require additional resources when compared to STTN+DVD. As it is difficult to estimate flow using off-the-shelf flow estimation methods in a recurrent model, it is not straightforward to find a natural extension of DVD to the recurrent architecture. For training, we use the dataset from [45] and the same optimizer (ADAM) using a constant learning rate (10^{-4}). The networks are trained with cropped patches (128x128) as suggested in [45], where the size of the minibatch is 8. For 700k

iterations, STTN+DVD minimizes the loss in (6), and RSTTN+DVD optimizes the loss in (7) for five recurrent steps. As shown by the quantitative comparison results given in Fig. 8 (d), the proposed STTN+DVD and RSTTN+DVD were able to restore two HD (1280x720) frames per second (i.e., 50 times faster than DVD), while improving deblurring quality in terms of PSNR by large margins. In Fig. 8 (e), we show that the proposed model removes spatially varying motion blur, and our RSTTN+DVD renders a significantly better deblurring result than the baseline model.

Comparison with OVD [47]. The online video deblurring (OVD) network runs in near real-time while giving state-of-the-art deblurring results. As illustrated in Fig. 9 (a), OVD is specialized for handling video frames in a recurrent manner. Nevertheless, our STTN+OVD, with a spatio-temporal transformer placed on top of OVD as shown in 9 (b), was able to further improve the deblurring quality. Note that since adding future frames plays a key role in OVD, our STTN+OVD model also takes a future frame $\mathbf{Y}^{t+1}$ as input of OVD.

For a fair comparison of these two networks, both OVD and our STTN+OVD networks are trained under the same conditions. We use 128x128 patches extracted from the video deblurring dataset [45] and use the ADAM optimizer with learning rate 10^{-4}. As these two networks are based on recurrent architectures, the gradient values are clipped to have magnitudes smaller than one to avoid the "exploding gradients" problem. We use a minibatch size of 8 and the networks are trained for five recurrent steps by minimizing the loss in (7) for 500k iterations. In Fig. 9 (c), quantitative comparison results are given. Our STTN+OVD model outperforms the original OVD, and the performance gap is around 0.25dB without the large computational overhead. We compare visual results in Fig. 9 (d), showing that STTN+DVD removes spatially varying blur at the occluded region of the moving bicycle significantly better than OVD, the current state-of-the-art.

5 Conclusions

We have proposed a novel spatio-temporal transformer network (STTN) which generalizes the spatial transformer network [13] while at the same time alleviating some of its limitations. Our STTN is composed of a spatio-temporal flow estimation module which calculates spatio-temporal flow in three dimensions from multiple image frames (or feature maps), and a spatio-temporal sampler which interpolates multiple inputs in spatio-temporal space. This way, the proposed model efficiently mitigates the problems of conventional flow estimation networks which suffer from unmatched regions, by exploiting multiple inputs at the same time rather than using a single target input. The superiority of the proposed model is demonstrated in a number of video restoration tasks, and we achieve state-of-the-art performance by simply adding the proposed module on top of conventional networks.

References

1. Liu, C., Sun, D.: A bayesian approach to adaptive video super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
2. Kim, T.H., Lee, K.M.: Generalized video deblurring for dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
3. Zhao, W., Sawhney, H.S.: Is super-resolution with optical flow feasible? In: Proceedings of the European Conference on Computer Vision (ECCV). (2002)
4. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
5. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. In: Association for the Advancement of Artificial Intelligence (AAAI). (2017)
8. Meister, S., Hur, J., Roth, S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In: AAAI. (2017)
9. Dosovitskiy, A., Springenberg, J.T., Brox, T.: Learning to generate chairs with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
10. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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 (CVPR). (2016)
11. Ahmadi, A., Patras, I.: Unsupervised convolutional neural networks for motion estimation. In: IEEE International Conference on Image Processing (ICIP). (2016)
12. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW). (2016)
13. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). (2015)
14. Zimmer, H., Bruhn, A., Weickert, J.: Optic flow in harmony. International Journal of Computer Vision (IJCV) **93**(3) (2011) 368–388
15. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)
16. Lee, K.J., Kwon, D., Yun, I.D., Lee, S.U.: Optical flow estimation with adaptive convolution kernel prior on discrete framework. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)
17. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)

18. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **34**(9) (2012) 1744–1757
19. Kim, T.H., Lee, H.S., Lee, K.M.: Optical flow via locally adaptive fusion of complementary data costs. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2013)
20. Volz, S., Bruhn, A., Valgaerts, L., Zimmer, H.: Modeling temporal coherence for optical flow. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2011)
21. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision (IJCV) **92**(1) (2011) 1–31
22. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2012)
23. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012)
24. Xu, J., Ranftl, R., Koltun, V.: Accurate optical flow via direct cost volume processing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
25. Güney, F., Geiger, A.: Deep discrete flow. In: Proceedings of the Asian Conference on Computer Vision (ACCV). (2016)
26. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Deepmatching: Hierarchical deformable dense matching. International Journal of Computer Vision (IJCV) **120**(3) (2016) 300–323
27. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
28. Mitzel, D., Pock, T., Schoenemann, T., Cremers, D.: Video super resolution using duality based tv-l1 optical flow. In: Joint Pattern Recognition Symposium. (2009)
29. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging **2**(2) (2016) 109–122
30. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
31. Makansi, O., Ilg, E., Brox, T.: End-to-end learning of video super-resolution with motion compensation. In: Proceedings of the German Conference on Pattern Recognition (GCPR). (2017)
32. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
33. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T.: Robust video super-resolution with learned temporal dynamics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
34. Sajjadi, M.S.M., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)

35. Cai, J.F., Ji, H., Liu, C., Shen, Z.: Blind motion deblurring using multiple images. Journal of computational physics **228**(14) (2009) 5057–5071
36. Zhang, H., Wipf, D., Zhang, Y.: Multi-image blind deblurring using a coupled adaptive sparse prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
37. Cho, S., Cho, H., Tai, Y.W., Lee, S.: Registration based non-uniform motion deblurring. Computer Graphics Forum **31**(7) (2012) 2183–2192
38. Zhang, H., Carin, L.: Multi-shot imaging: joint alignment, deblurring and resolution-enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
39. Zhang, H., Yang, J.: Intra-frame deblurring by leveraging inter-frame camera motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
40. Li, Y., Kang, S.B., Joshi, N., Seitz, S.M., Huttenlocher, D.P.: Generating sharp panoramas from motion-blurred videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)
41. Wulff, J., Black, M.J.: Modeling blurred video with layers. In: Proceedings of the European Conference on Computer Vision (ECCV). (2014)
42. Kim, T.H., Nah, S., Lee, K.M.: Dynamic video deblurring using a locally adaptive linear blur model. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2018)
43. Kim, T.H., Lee, K.M.: Segmentation-free dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
44. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
45. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
46. Wieschollek, P., Hirsch, M., Schölkopf, B., Lensch, H.P.: Learning blind motion deblurring. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
47. Kim, T.H., Lee, K.M., Schölkopf, B., Hirsch, M.: Online video deblurring via dynamic temporal blending network. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (MICCAI), Springer (2015) 234–241
49. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning (2015)
50. Sajjadi, M.S.M., Schölkopf, B., Hirsch, M.: EnhanceNet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
51. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
52. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: Tv-l1 optical flow estimation. Image Processing On Line (IPOL) (2013)