Flow-Guided Transformer for Video Inpainting

Kaidong Zhang[1], Jingjing Fu[2](✉), and Dong Liu[1]

[1]University of Science and Technology of China [2]Microsoft Research Asia
richu@mail.ustc.edu.cn, jifu@microsoft.com, dongeliu@ustc.edu.cn

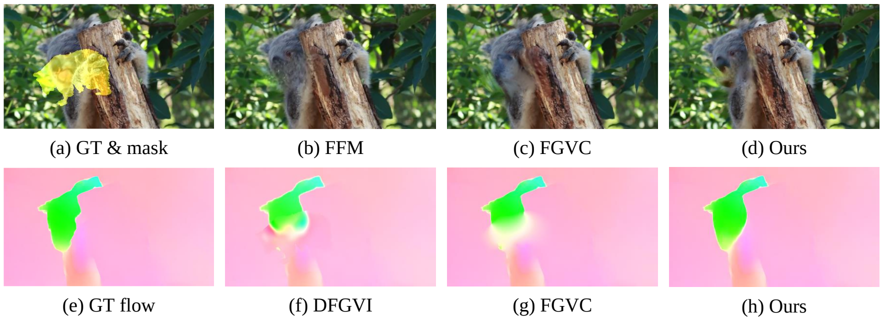

(a) GT & mask (b) FFM (c) FGVC (d) Ours
(e) GT flow (f) DFGVI (g) FGVC (h) Ours

Fig. 1: Performance comparison on frame synthesis (top row) and flow completion (bottom row) between our method and some state-of-the-art baselines [50,14,29]. Our method achieves significant performance improvement against the compared baselines and obtains more coherent results.

Abstract. We propose a flow-guided transformer, which innovatively leverage the motion discrepancy exposed by optical flows to instruct the attention retrieval in transformer for high fidelity video inpainting. More specially, we design a novel flow completion network to complete the corrupted flows by exploiting the relevant flow features in a local temporal window. With the completed flows, we propagate the content across video frames, and adopt the flow-guided transformer to synthesize the rest corrupted regions. We decouple transformers along temporal and spatial dimension, so that we can easily integrate the locally relevant completed flows to instruct spatial attention only. Furthermore, we design a flow-reweight module to precisely control the impact of completed flows on each spatial transformer. For the sake of efficiency, we introduce window partition strategy to both spatial and temporal transformers. Especially in spatial transformer, we design a dual perspective spatial MHSA, which integrates the global tokens to the window-based attention. Extensive experiments demonstrate the effectiveness of the proposed method qualitatively and quantitatively. Codes are available at https://github.com/hitachinsk/FGT.

Keywords: Video inpainting, Optical flow, Transformer

1 Introduction

Video inpainting aims at filling the corrupted regions in a video with reasonable and spatiotemporally coherent content [2]. Its application includes but not limited to watermark removal [34], object removal [15], video retargeting [22], and video stabilization [31]. High-quality video inpainting is challenging because it requires spatiotemporal consistency of the restored video. Directly applying image inpainting methods [37,20,27,53,54,33,51,38,26] is sub-optimal, because they mainly refer the content within one frame but fail to utilize the complementary content across the whole video.

Transformer [43] has sparked the computer vision community. Its outstanding long-range modeling capacity makes it naturally suitable for video inpainting, as video inpainting relies on the content propagation across frames spatiotemporally to fill the corrupted regions with high fidelity. Previous works [55,29,28] modify transformer for video inpainting task, and achieve unprecedented performance. However, these works still suffer from inaccurate attention retrieval. They mainly utilize the appearance features in transformer, but ignore the object integrity exposed by the motion fields, which indicates the relevant regions.

Recently, several works [50,14,57] propose to complete optical flows for video inpainting. As discussed in DFGVI [50], optical flows are much easier to complete because they contain far less complex patterns than frames. Since the relative motion magnitude between foreground objects and background are different, the contents with similar motion pattern are more likely to be relevant. Therefore, the motion discrepancy of optical flows can serve as a strong instructor to guide the attention retrieval for more relevant content. Inspired by this, we propose a novel flow-guided transformer to synthesize the corrupted regions with the motion guidance from completed flows. Our method contains two parts: the first is a flow completion network designed to complete the corrupted flows, and the second is the flow-guided transformer proposed to synthesize the corrupted frames under the guidance of the completed flows.

During flow completion, we observe that the flows in a local temporal window are more correlated than the distant ones, because motion fields are likely to be maintained in a short temporal window. Therefore, we propose to exploit the correlation of complementary features of optical flows in a local temporal window, which is different from the simply stacking strategy in DFGVI [50] and the single flow completion method in FGVC [14]. We integrate spatial-temporal decoupled P3D blocks [39] to a simple U-Net [40], which completes the target flow based on the local reference flows. Furthermore, we propose a novel edge loss to supervise the completion quality in the edge regions without introducing additional computation cost during inference. Compared with previous counterparts [50,14], our method can complete more accurate flows.

Under guidance of the completed optical flows, we propagate the content from the valid regions to the corrupted regions, and then synthesize the rest corrupted content in the video frames with the flow-guided transformer. Following previous transformer-based video inpainting methods [55,29,28], we sample video frames from the whole video and inpaint these frames simultaneously. Given that the optical flows are locally correlated, we decouple the spatial and temporal attention in transformer and only integrate optical flows into spatial transformers. In temporal transformer, we perform multi-head self-attention (MHSA) spatiotemporally, while in spatial transformer, we only perform MHSA within the tokens coming from the same frame. Considering that the completed flows are not perfect and the content with different appearance may have similar motion patterns, we propose a novel flow-reweight module to control the impact of flow tokens based on the interaction between frame and flow tokens adaptively.

To improve the efficiency of our transformer, we introduce window partition strategy [30,52,9] in the flow-guided transformer. In temporal transformer, as the temporal offset between distant frames may be large, small temporal window size cannot include abundant temporal relevant tokens. As a result, we perform MHSA in a large window to exploit rich spatiotemporal tokens. In spatial transformer, after flow guidance integration, we restrict the attention within a smaller window based on local smoothness prior of natural images. However, simple window attention ignores the possible correlated content at the distant location. To relieve such problem, we extract the tokens from the whole token map globally and integrate these global tokens to the key and value. In such manner, the queries can not only retrieve the fine-grained local tokens, but also attend to the global content. We refer this design as dual perspective spatial MHSA.

We conduct extensive experiments to validate the effectiveness of different components of our method. As shown in Fig. 1, our method remarkably outperforms previous baselines in terms of visualization results on frame synthesis and flow completion. In summary, our contributions are:

- We propose a flow-guided transformer to integrate the completed optical flow into the transformer for more accurate attention retrieval in video inpainting.
- We design a novel flow completion network with local flow features exploitation, which outperforms previous methods significantly.
- We introduce window partition strategy in the video inpainting transformer and propose the dual perspective spatial MHSA to enrich the local window attention with global content.

2 Related Work

Traditional methods. Traditional video inpainting methods [2,15,31,12,16,34] explore the geometry relationship (e.g. homography or optical flows) between the corrupted regions of the target frames and the valid regions of the reference frames for content synthesis with high fidelity. Huang *et al.*[19] design a set of energy equation to optimize optical flow reconstruction and frame synthesis interactively and achieve unprecedented video inpainting quality.

Deep learning based methods. Deep learning based methods can be divided into two categories, the first one [50,14] aims to complete the missing optical flows to capture the motion correlation between the valid regions and the corrupted regions. Our method also includes the flow completion component, but we only

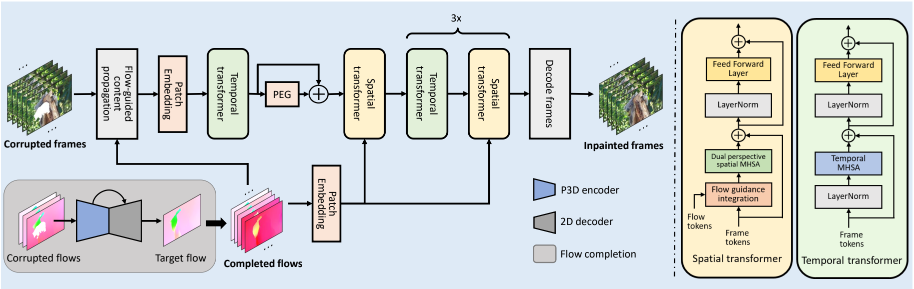


Fig. 2: Our method consists of two steps. Firstly, we adopt the **L**ocal **A**ggregation **F**low **C**ompletion network (LAFC) to complete the corrupted target flow, and then propagate the content among the video frames with the completed flows. Secondly, we synthesize the rest corrupted regions with **F**low-**G**uided **T**ransformer (FGT). PEG: Position embedding generator.

exploit the complementary flow features in a local window for more efficient and accurate flow completion.

The second category targets on directly synthesizing the corrupted regions from video frames. Some works adopt 3D CNN [44,7] or channel shift [8,59,21] to model the complementary features between local frames. Several methods integrate recurrent [22,25] or attention [24,35] mechanism into CNN-based networks to expand the temporal receptive field. Inspired by the spatiotemporal redundancy in videos, Zhang *et al.*[56] and Ouyang *et al.*[36] adopt internal learning to perform long range propagation for video inpainting. Currently, Zeng *et al.*[55] and Liu *et al.*[29,28] design specific transformer [43] to retrieve similar features in a considerable temporal receptive field for high-quality video inpainting. Our method is also built upon transformer, but differently we improve the attention retrieval accuracy with the completed flows.

Transformer in vision. Due to the outstanding long range feature capture ability, transformer [43] has been introduced to various computer vision tasks, such as basic architecture design [30,52,9], image classification [11,3,13,48], object detection [6,32], action detection [46], segmentation [45], etc. We revisit the design of transformer in video inpainting and propose several strategies to improve efficiency while maintaining competitive performance, including spatial-temporal decomposition and the combination of local and global tokens.

3 Method

3.1 Problem formulation

Given a corrupted video sequence $X := \{X_1, ..., X_T\}$, whose corrupted regions are annotated by the corresponding mask sequence $M := \{M_1, ..., M_T\}$. T is video length. Our goal is to generate the inpainted video sequence $\hat{Y} := \{\hat{Y}_1, ..., \hat{Y}_T\}$

and maintain the spatiotemporal coherence between our result and the ground truth video sequence $Y :=\{Y_1, ..., Y_T\}$.

3.2 Network overview

As shown in Fig. 2, our network consists of a **L**ocal **A**ggregation **F**low **C**ompletion network (LAFC) for flow completion and a **F**low-**G**uided **T**ransformer (FGT) to synthesize the corrupted regions. For a given masked video sequence X, we extract its forward and backward optical flows $\tilde{F}_f$ and $\tilde{F}_b$ and utilize LAFC to complete each optical flow with its local references. Based on completed flows, we propagate the content across video frames. As for the rest corrupted regions, we adopt FGT to synthesize them.

3.3 Local aggregation flow completion network

Local flow aggregation. The motion direction and velocity of objects vary overtime, and the correlation between distant optical flows may be degraded severely. Fortunately, the variance of motion in short time is a gradual process, which means optical flows in a short temporal window are highly correlated, and they are reliable references for more accurate flow completion.

3D convolution block [42] is suitable to capture the local relevant content spatiotemporally. However, the parameter and computation overhead of 3D convolution block are large, which increases the difficulty for network optimization. Considering efficiency, we adopt P3D block [39] instead to decouple the local flow feature aggregation along temporal and spatial dimension. We insert P3D blocks to the encoder of LAFC and add skip connection [40] to exploit the local correlation between flows. Considering that LAFC completes forward and backward optical flows in the same manner, we denote both F_f and F_b as F for simplicity. Given a corrupted flow sequence, we utilize Laplacian filling to obtain the initialized flows $\tilde{F}$=$\{\tilde{F}_{t-ni}, ..., \tilde{F}_t, ..., \tilde{F}_{t+ni}\}$, where $\tilde{F}_t$ is the target corrupted flow, i is the temporal interval between consecutive flows, and the length of the flow sequence is $2n + 1$. The initialized flow sequence $\tilde{F}$ are fed to the LAFC to complete the target flow $\tilde{F}_t$. We denote the input of m-th P3D block as $\tilde{f}^m$, and the output as $\tilde{f}^{m+1}$. The local feature aggregation process can be formulated as.

$$\tilde{f}^{m+1} = \mathrm{TC}(\mathrm{SC}(\tilde{f}^m)) + \tilde{f}^m \tag{1}$$

Where TC represents 1D temporal convolution, and SC is the 2D spatial convolution. We keep the temporal resolution unchanged except the final P3D block in the encoder and the P3D blocks inserted in the skip connection. In these blocks, we shrink the temporal resolution of the flow sequence to obtain the aggregated flow features of the target flow. Finally, a 2D decoder is utilized to obtain the completed target optical flow $\hat{F}_t$.

Edge loss. In general, flow fields are piece-wise smooth, which means the flow gradients are considerable small except motion boundaries [14]. The edges in flow maps inherently contain crucial salient features that may benefit the reconstruction of object boundaries. Nevertheless, the flow completion in edge regions is a tough task, as there is no specified guidance to edge recovery. Therefore, we design a novel edge loss in LAFC to supervise the completion quality in edge regions of $\hat{F}_t$ explicitly, which can improve the flow completion quality without introducing additional computation overhead during inference.

For the completed target flow $\hat{F}_t$, we extract the edges with a small projection network P_e and calculate the binary cross entropy loss with the edges that extracted from the ground truth F_t with Canny edge detector [5].

$$L_e = \mathrm{BCE}(\mathrm{Canny}(F_t), P_e(\hat{F}_t)) \tag{2}$$

where L_e is the edge loss. We utilize four convolution layers with residual connection [17] to formulate P_e.

Loss function. We adopt L1 loss to penalize $\hat{F}_t$ in the corrupted and the valid regions, respectively. To improve the smoothness of $\hat{F}_t$, we impose first and second order smoothness loss to $\hat{F}_t$.

What's more, we also warp the corresponding ground truth frames with $\hat{F}_t$ to penalize the regions with large warp error. We adopt the L1 loss to supervise the warping quality, and expel the occlusion regions with forward-backward consistency check of ground truth optical flows for more accurate loss calculation. The loss function of LAFC is the combination of the loss terms discussed above, and the detailed formulas are provided in the supplementary material.

3.4 Flow-guided transformer for video inpainting

After flow completion, we propagate the content from valid regions to corrupted regions throughout the whole video to fill-in the corrupted regions that can be connected with the valid regions. The rest corrupted regions are filled with our designed flow-guided transformer (FGT). FGT takes multiple corrupted frames into consideration and synthesize these frames simultaneously. Since the motion discrepancy of completed optical flows to some extent reveals the shape and location of foreground objects and background, we integrate such information to FGT to indicate the relevant regions inside a single frame. Due to the degraded correlation between distant optical flows, the traditional all-pair interaction between tokens from distant frames may not be suitable for flow guidance integration. Therefore, we decouple MHSA along the temporal and spatial dimension, and we only integrate the flow content to the spatial MHSA.

In both spatial and temporal transformer blocks, we introduce specific designs for efficiency and performance balance. In temporal transformer, we adopt large window to compensate the reference offset between distant frames. In spatial transformer, we divide each token map into small window based on the local smoothness prior of natural images, and supply the key and value with the condensed global tokens to perform spatial MHSA in dual perspective from local and global views.

As shown in Fig. 2, given the frame sequence $\hat{X}$ after flow-guided content propagation, we crop $\hat{X}$ and completed flows into patches and project them to

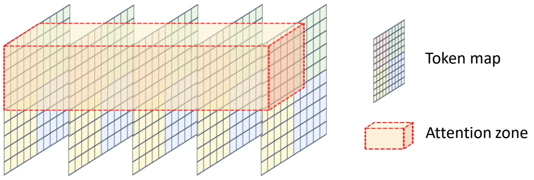


Fig. 3: The temporal MHSA in the temporal transformer. We split non-overlapped large windows (zones) for each token map, and perform MHSA inside the cube formed by the corresponding position in each token map. The windows are shown with different colors. In this figure, we illustrate the 2×2 zone.

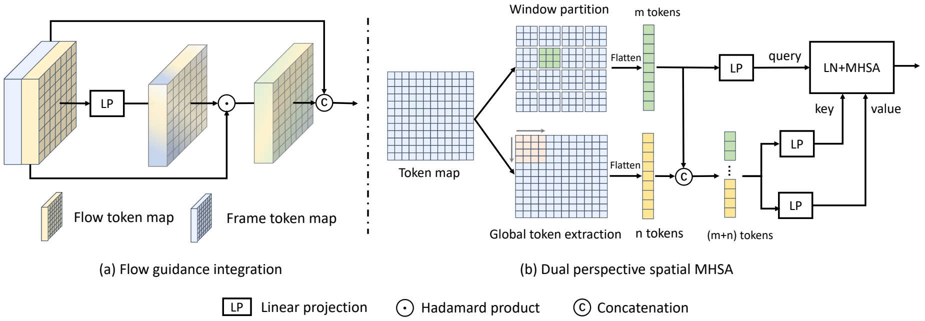


Fig. 4: Illustration of flow guidance integration and dual perspective spatial MHSA in the spatial transformer.

frame tokens TI with an encoder. The completed flows are also projected to flow tokens TF. We refer such process as "patch embedding". We design interleaved temporal and spatial transformer blocks to process TI, and enrich the frame tokens with TF in each spatial transformer block before dual perspective spatial MHSA. As for positional embedding, we follow CVPT [10] to adopt depth-wise convolution [18] after the first transformer block for video inpainting in flexible resolutions, while the pre-defined trainable positional embedding of previous works [29] can only process the videos at certain resolution.

Temporal transformer. In temporal transformer, attention retrieval is performed to the tokens across different frames. Since the content shifts along temporal dimension, it is reasonable to apply large size window to compensate the reference offset. Liu *et al.*[28] also demonstrate the all-pair attention strategy is unnecessary in video inpainting. Therefore, we divide each token maps in TI into non-overlap cubes with large window size (denoted as "zone") along height and width dimension and perform MHSA within the cubes, as shown in Fig. 3.

Flow guidance integration. The motion discrepancy between different objects and background exposed by optical flows indicates the content relationship. The tokens with similar motion magnitude are more likely to be relevant. Therefore, we utilize the completed optical flows to guide the attention process in FGT.

As discussed in Sec. 3.3, optical flows are locally correlated. Therefore, we only exploit the optical flows in the spatial transformer. A straightforward way is to concatenate TI and TF along channel dimension directly before spatial MHSA. However, there are two problems in this way. First, the completed flows are not perfect. The distorted flows may mislead the judgement about relevant regions. Second, the appearance patches may vary a lot within objects, while the corresponding motion patterns may still be similar, which is likely to confuse the attention retrieval process. In order to ease these problems, we propose a flow-reweight module to control the impact of flow tokens TF with respect to the interaction between TF and TI, as shown in Fig. 4(a). We formulate the flow-reweight module as.

$$\hat{TF}_t^j = TF_t^j \odot \mathrm{MLP}(\mathrm{Concat}(TI_t^j, TF_t^j)) \tag{3}$$

where Concat is the concatenation operation. MLP represents the MLP layers, and $\hat{TF}_t^j$ represents the t-th reweighted flow token map in j-th spatial transformer. Finally, we concatenate $\hat{TF}_t^j$ and TI_t^j to obtain the flow-enhanced tokens TK_t^j to enhance spatial MHSA.

Dual perspective spatial MHSA. We introduce window partition to spatial MHSA for efficiency. According to the local smoothness prior of natural images, the tokens are more correlated to their neighbors. Hence, we adopt relative small window size in spatial transformer. Given the t-th frame token map processed by j-th transformer $TK_t^j \in \mathbb{R}^{H \times W \times C}$, where H, W and C represent the height, width and channel size. Window partition divides TK_t^j into several $h \times w$ non-overlapped windows, and MHSA is performed inside each window, respectively. However, if the window contains numerous tokens projected from corrupted regions, the attention accuracy would be deteriorated due to the lack of valid content. Therefore, we integrate global tokens to spatial MHSA. We adopt depth-wise convolution [18] to condense TK_t^j to global tokens, and supply them to each window. Given the kernel size k and downsampling rate s (also known as stride), the global tokens are generated as.

$$TG_t^j = \mathrm{DC}(TK_t^j, k, s) \tag{4}$$

where TG_t^j represents the extracted global tokens and DC is the depth-wise convolution. The query $Q_t^j(d)$, key $K_t^j(d)$ and value $V_t^j(d)$ of the d-th window in TK_t^j are generated as.

$$\begin{aligned} Q_t^j(d) &= \mathrm{MLP}(\mathrm{LN}(TK_t^j(d))) \\ K_t^j(d) &= \mathrm{MLP}(\mathrm{LN}(\mathrm{Concat}(TK_t^j(d), TG_t^j))) \\ V_t^j(d) &= \mathrm{MLP}(\mathrm{LN}(\mathrm{Concat}(TK_t^j(d), TG_t^j))) \end{aligned} \tag{5}$$

where $TK_t^j(d)$ represents the d-th window in TK_t^j, and LN is layer normalization [1]. After we obtain $Q_t^j(d)$, $K_t^j(d)$ and $V_t^j(d)$, we apply spatial MHSA to process them. The dual perspective spatial MHSA is illustrated in Fig. 4(b).

Note that the global tokens are shared by all the windows. In each spatial transformer, if we adopt all-pair attention retrieval for MHSA, each token will retrieve $H \times W$ tokens. While the token number for retrieval in our dual perspective spatial MHSA is $(\lceil \frac{H}{s} \rceil \times \lceil \frac{W}{s} \rceil + h \times w)$. It is easy to derive that when $s \geq \lceil \sqrt{\frac{HW}{HW-hw}} \rceil$, the referenced token number will be smaller than the token number in all-pair attention retrieval.

Recently, focal transformer [52] has also adopted the combination of local and global attention in transformer. Compared with [52], our method decouples the global token size and the window shape, which is more flexible than the sub-window pooling strategy in focal transformer.

Loss function We adopt the reconstruction loss in the corrupted and the valid regions together with the T-Patch GAN loss [7] to supervise the training process. We use hinge loss as the adversarial loss. We provide the detailed loss formulas in the supplementary material.

4 Experiments

4.1 Settings

We adopt Youtube-VOS [49] and DAVIS [4] datasets for evaluation. Youtube-VOS contains 4453 videos and DAVIS contains 150 videos. We adopt the training set of Youtube-VOS to train our networks. As for Youtube-VOS, we evaluate the trained models on its testset. Since DAVIS contains densely annotated masks on its training set, we adopt its training set to evaluate our method.

Following the previous work [14], we choose PSNR, SSIM [47] and LPIPS [58] as our evaluation metrics. Meanwhile, we adopt end-point-error (EPE) to evaluate the flow completion quality. We compare our method with state-of-the-art baselines, including VINet [22], DFGVI [50], CPN [24], OPN [35], 3DGC [7], STTN [55], FGVC [14], TSAM [59], DSTT [28] and FFM [29].

4.2 Implementation details

In our experiments, We utilize RAFT [41] to extract optical flows. In flow completion network, the flow interval and input flow number are both set to 3. The flow locating in middle of the local temporal window is treated as the target flow for completion. We adopt gradient propagation [14] as our flow-guided content propagation strategy, and the detailed procedure will be provided in the supplementary material. As for FGT, we keep the patch embedding method the same as FFM [29] for fair comparisons. We utilize the forward optical flows in the flow guidance integration module. FGT adopts 8 transformer blocks in total (4 temporal and 4 spatial transformer blocks). In the temporal transformer, we adopt 2×2 zone division for temporal MHSA. In the spatial transformer, the downsampling rate of the global token is 4, while the window size is 8. We adopt Adam optimizer [23] to train our networks. The training iteration is 280K for LAFC and 500K for FGT. The initial learning rate is $1e$-4, which is divided by

Table 1: Quantitative results on the Youtube-VOS and DAVIS datasets. The best and second best numbers for each metric are indicated by red and blue fonts, respectively. ↓ means lower is better, while ↑ means higher is better. "FGT" represents we adopt our proposed flow-guided transformer to fill all the pixels in the corrupted regions without flow-guided content propagation.

Method	Youtube-VOS			DAVIS					
				square			object		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VINet [22]	29.83	0.955	0.047	28.32	0.943	0.049	28.47	0.922	0.083
DFGVI [50]	32.05	0.965	0.038	29.75	0.959	0.037	30.28	0.925	0.052
CPN [24]	32.17	0.963	0.040	30.20	0.953	0.049	31.59	0.933	0.058
OPN [35]	32.66	0.965	0.039	31.15	0.958	0.044	32.40	0.944	0.041
3DGC [7]	30.22	0.961	0.041	28.19	0.944	0.049	31.69	0.940	0.054
STTN [55]	32.49	0.964	0.040	30.54	0.954	0.047	32.83	0.943	0.052
TSAM [59]	31.62	0.962	0.031	29.73	0.951	0.036	31.50	0.934	0.048
DSTT [28]	33.53	0.969	0.031	31.61	0.960	0.037	33.39	0.945	0.050
FFM [29]	33.73	0.970	0.030	31.87	0.965	0.034	34.19	0.951	0.045
FGT	34.04	0.971	0.028	32.60	0.965	0.032	34.30	0.953	0.040
FGVC [14]	33.94	0.972	0.026	32.14	0.967	0.030	33.91	0.955	0.036
Ours	34.53	0.976	0.024	33.41	0.974	0.023	34.96	0.966	0.029

10 after 120K iterations for LAFC and 300K iterations for FGT. For ablation studies, following FFM [29], we train FGT for 250K iterations, and the learning rate is divided by 10 after 200K iterations. We perform ablation studies on DAVIS dataset.

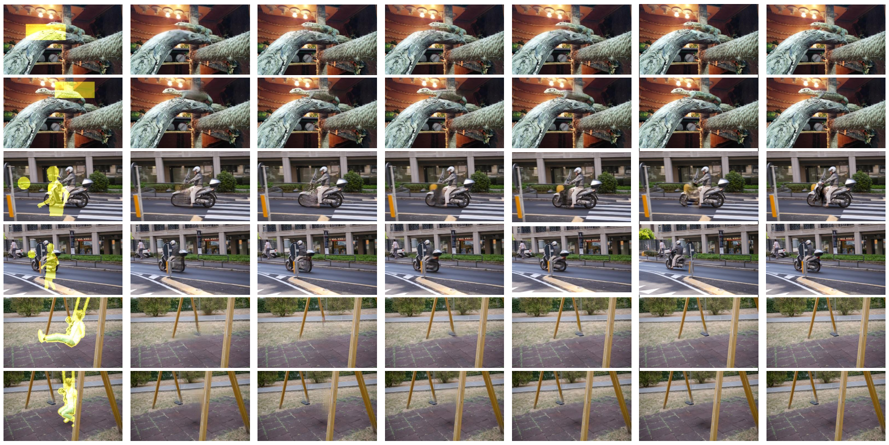
4.3 Quantitative evaluation

During inference, the resolution of videos is set to 432×256. We generate square masksets with continuous motion trace for Youtube-VOS and DAVIS datasets. The average size of the masks in the square maskset is $\frac{1}{16}$ of the whole frame. We shuffle DAVIS object maskset randomly and corrupt frames with these masks to evaluate video inpainting performance upon object masks. For fair comparisons among flow-based video inpainting methods, we utilize the same optical flow extractor for DFGVI [50], FGVC [14] and our method.

We report the quantitative evaluation results of our method and other baselines in Tab. 1. Our method outperforms previous baselines by a significant margin on all three metrics, which means the restored videos from our method enjoy less distortion and better perceptual quality than previous counterparts. What's more, if we fill the corrupted region purely with FGT, we can still outperform previous transformer-based video inpainting baselines [55,29,28].

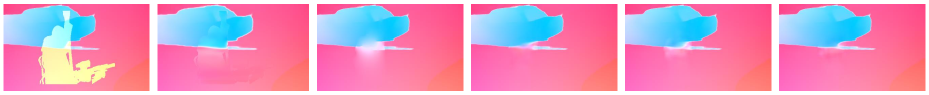
4.4 Qualitative comparisons

We compare the qualitative results between our method and five recent baselines [55,14,59,28,29] under the square mask, object mask and object removal settings. The results are shown in Fig. 5. Compared with these baselines, our



(a) Input (b) STTN [55] (c) TSAM [59] (d) DSTT [28] (e) FFM [29] (f) FGVC [14] (g) Ours

Fig. 5: Qualitative comparison between our method and some recent baselines [55,14,59,28,29]. From top to bottom, every two rows display inpainting results of square mask set, object mask set, and object removal, respectively.



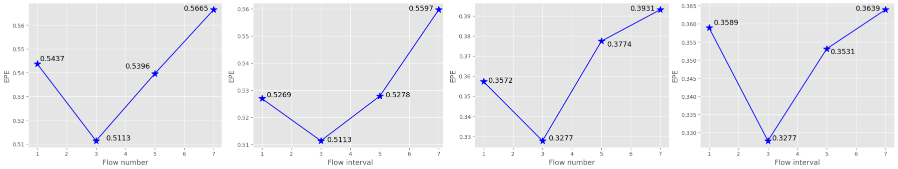
(a) Input (b) DFGVI (c) FGVC (d) S (d) LA (d) LA + L_e

Fig. 6: Comparison of flow results between DFGVI [50], FGVC [14], and several variants of our method. S: single flow completion, LA: Flow completion with local aggregation, L_e: Edge loss.

method enjoys outstanding visual quality. Our method can complete more accurate optical flows, which describes the motion trajectory with high fidelity. Therefore, our method enjoys less distortion in the content propagation stage than FGVC [14]. What's more, the completed optical flows provide accurate object clusters. Such information leads to more accurate attention retrieval and naturally produce better visual quality. We will provide more video inpainting results in the supplementary materials.

4.5 Ablation Studies

Model analysis. We compare our method with (1) FGVC and (2) FGVC+ FGT to justify the design of our method over flow completion and image inpainting baseline [53]. The results in Tab. 2(a) demonstrate the effectiveness of our method in both flow completion and frame synthesis. In Tab. 2(b), we compare FGT with different transformer baselines. Since FLOPs in video inpainting is related to the number of frames processed simultaneously, we assume



(a) Square masks

(a) Object masks

Fig. 7: EPE results with varying flow number (when flow interval is 3) or varying flow interval (when flow number is 3) on both square and object mask sets.

Table 2: **Model analysis** We report the analysis of the method variants and the comparison of the efficiency between FGT and other baselines.

(a) **Analysis about method variants.**

Method	square			object		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
FGVC [14]	32.14	0.967	0.030	33.91	0.955	0.034
FGVC+FGT	32.49	0.968	0.027	34.58	0.956	0.031
LAFC+FGT	**33.41**	**0.974**	**0.023**	**34.96**	**0.966**	**0.029**

(b) **Efficiency analysis.**

Method	FLOPs(per frame)	Params	Speed
STTN [55]	477.91G	16.56M	0.22s
FFM [29]	579.82G	36.59M	0.30s
FGT(all-pair)	703.22G	42.31M	-
FGT	455.91G	42.31M	0.39s

the processed frame number is 20, which is a common practice in STTN [55] and FFM [29]. "FGT(all-pair)" means we adopt all-pair attention in FGT, which consumes much more computation overhead compared with FGT. If we adopt flow-guided content propagation, we can obtain better video inpainting quality, but the speed will degrade to 2.11s/frame, which indicates the performance-efficiency trade-off in our method. We provide detailed run-time analysis in the supplementary material.

Local flow aggregation and edge loss for flow completion. We report the end-point-error (EPE) of single flow completion (replace the P3D blocks with vanilla convolution blocks), local aggregation for flow completion without and with edge loss, together with two baselines [50,14] in Tab. 3. With the introduction of local aggregation and edge loss, our method achieves substantial improvement. The subjective results are shown in Fig. 6. With local aggregation, our method can exploit the complementary flow features in a local temporal window, which is beneficial to complete accurate flow shape. With edge loss, our method can synthesize optical flows with clearer motion boundaries. Finally, we report the influence of flow number and flow interval w.r.t. EPE in Fig. 7. When the flow number or interval is too small, the target flow cannot utilize abundant references for accurate flow completion, which undermines the performance. When the flow number or interval is large, the flow completion performance deteriorates gradually, which reveals the distant flows contribute less to flow completion relative to local flows.

Flow guidance integration and dual perspective spatial MHSA. In this part, we adopt the transformer to synthesize all the pixels in the corrupted re-

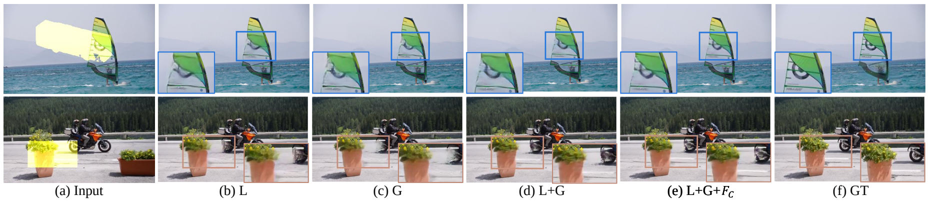

(a) Input (b) L (c) G (d) L+G (e) L+G+F_c (f) GT

Fig. 8: Qualitative comparison of different components in the dual perspective transformer. L: Local window attention. G: Global tokens. F_c: Flow guidance with flow-reweight module.

Table 3: Ablation study about flow completion. S: single flow completion, LA: Flow completion with local aggregation, L_e: Edge loss.

Maskset	EPE↓				
	DFGVI [50]	FGVC [14]	S	LA	LA + L_e
square	1.161	0.633	0.546	0.524	0.511
object	1.053	0.491	0.359	0.338	0.328

gions for fair comparisons across different settings. We evaluate the effectiveness of the dual perspective tokens, the completed flow guidance and the flow-reweight module in spatial MHSA, and report the corresponding results in Tab. 4.

The quantitative results demonstrate the effectiveness of our proposed method. Compared with attention with only local or global tokens, the combination of these two perspective tokens achieves significant performance boost. With the introduction of the completed flow tokens and the flow-reweight module, the performance of our model boosts further. When we remove the flow-reweight module, the performance degrades, which demonstrates the necessity to introduce flow guidance and control its impact during attention retrieval.

The qualitative comparisons between different components in our flow-guided transformer are shown in Fig. 8. We can observe the substantial improved visual quality on dual perspective attention and the introduction of flow guidance. The combination of global and local tokens enlarges the attention retrieval space while maintaining the local smoothness simultaneously. As for flow guidance, we visualize the local and global attention maps in Fig. 9. The red square in Fig. 9(a) indicates the query token. With flow guidance, our transformer tends to query the tokens with similar motion pattern (e.g. tokens in car region), which leads to clearer object boundary for video inpainting in higher quality.

5 Conclusion

In this work, we propose a flow-guided transformer for video inpainting, which introduces a novel way to leverage the motion discrepancy from optical flows

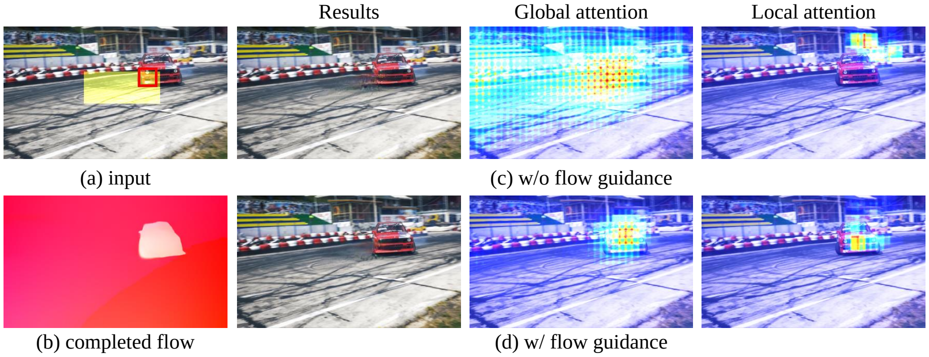


Fig. 9: Attention map visualization of our transformer model with/without the flow guidance. The red square in (a) indicates the location of the chosen query token for visualization.

Table 4: Ablation study about the spatial transformer. W: Local window partition. G: Global tokens. F_C: Completed flow tokens. RF: Flow-reweight module.

W	G	F_C	RF	square			object		
				PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
✓	-	-	-	31.37	0.957	0.038	32.98	0.945	0.051
-	✓	-	-	31.42	0.958	0.040	33.10	0.945	0.050
✓	✓	-	-	31.62	0.959	0.038	33.25	0.946	0.048
✓	✓	✓	-	31.54	0.958	0.039	33.12	0.945	0.049
✓	✓	✓	✓	31.87	0.961	0.036	33.52	0.947	0.045

to instruct the attention retrieval in transformer. We decouple the attention module along spatial and temporal dimension to facilitate the integration of the completed flows. We propose the flow-reweight module to control the impact of the flows in the attention retrieval process. What's more, in both temporal and spatial transformer blocks, we design specific window partition strategy for better efficiency while maintaining the competitive performance. Besides the proposed flow-guided transformer, We design a flow completion network to exploit the complementary features of the optical flows in a local temporal window, and introduce edge loss to supervise the reconstruction of flows for clear motion boundaries. The high-quality completed flows benefit the content propagation and flow-guided transformer. Extensive experiments demonstrate the effectiveness of our proposed method.

Acknowledgment. This work was supported by the Natural Science Foundation of China under Grants 62036005, 62022075, and 62021001, and by the Fundamental Research Funds for the Central Universities under Contract No. WK3490000006.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016) 8
2. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: CVPR. vol. 1, pp. 355–362 (2001) 2, 3
3. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: ICCV. pp. 10231–10241 (October 2021) 4
4. Caelles, S., Montes, A., Maninis, K.K., Chen, Y., Gool, L.V., Perazzi, F., Pont-Tuset, J.: The 2018 DAVIS challenge on video object segmentation. arXiv preprint arXiv:1803.00557 (2018) 9
5. Canny, J.: A computational approach to edge detection. TPAMI **PAMI-8**(6), 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851 6
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020) 4
7. Chang, Y.L., Liu, Z.Y., Lee, K.Y., Hsu, W.: Free-form video inpainting with 3D gated convolution and temporal PatchGAN. In: ICCV. pp. 9066–9075 (2019) 4, 9, 10
8. Chang, Y.L., Liu, Z.Y., Lee, K.Y., Hsu, W.: Learnable gated temporal shift module for deep video inpainting. In: BMVC (2019) 4
9. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. In: NeurIPS (2021) 3, 4
10. Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C.: Conditional positional encodings for vision transformers. Arxiv preprint 2102.10882 (2021), `https://arxiv.org/pdf/2102.10882.pdf` 7
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021), `https://openreview.net/forum?id=YicbFdNTTy` 4
12. Ebdelli, M., Le Meur, O., Guillemot, C.: Video inpainting with short-term windows: Application to object removal and error concealment. TIP **24**(10), 3034–3047 (2015). https://doi.org/10.1109/TIP.2015.2437193 3
13. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: ICCV. pp. 6824–6835 (October 2021) 4
14. Gao, C., Saraf, A., Huang, J.B., Kopf, J.: Flow-edge guided video completion. In: ECCV. pp. 713–729 (2020) 1, 2, 3, 5, 9, 10, 11, 12, 13
15. Granados, M., Tompkin, J., Kim, K., Grau, O., Kautz, J., Theobalt, C.: How not to be seen — object removal from videos of crowded scenes. Comput. Graph. Forum **31**(2pt1), 219–228 (may 2012). https://doi.org/10.1111/j.1467-8659.2012.03000.x, `https://doi.org/10.1111/j.1467-8659.2012.03000.x` 2, 3
16. Granados, M., Kim, K.I., Tompkin, J., Kautz, J., Theobalt, C.: Background inpainting for videos with dynamic objects and a free-moving camera. In: ECCV. pp. 682–695 (2012) 3
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 6
18. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017) 7, 8

19. Huang, J.B., Kang, S.B., Ahuja, N., Kopf, J.: Temporally coherent completion of dynamic video. TOG **35**(6), 196:1–11 (2016) 3
20. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. TOG **36**(4), 107:1–14 (2017) 2
21. Ke, L., Tai, Y.W., Tang, C.K.: Occlusion-aware video object inpainting. In: ICCV (2021) 4
22. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Deep video inpainting. In: CVPR. pp. 5792–5801 (2019) 2, 4, 9, 10
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2014) 9
24. Lee, S., Oh, S.W., Won, D., Kim, S.J.: Copy-and-paste networks for deep video inpainting. In: ICCV. pp. 4413–4421 (2019) 4, 9, 10
25. Li, A., Zhao, S., Ma, X., Gong, M., Qi, J., Zhang, R., Tao, D., Kotagiri, R.: Short-term and long-term context aggregation network for video inpainting. In: ECCV. p. 728–743 (2020) 4
26. Liao, L., Xiao, J., Wang, Z., Lin, C.W., Satoh, S.: Image inpainting guided by coherence priors of semantics and textures. In: CVPR. pp. 6539–6548 (June 2021) 2
27. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: ECCV. pp. 85–100 (2018) 2
28. Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., Li, H.: Decoupled spatial-temporal transformer for video inpainting (2021) 2, 4, 7, 9, 10, 11
29. Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., Li, H.: Fuseformer: Fusing fine-grained information in transformers for video inpainting. In: ICCV (2021) 1, 2, 4, 7, 9, 10, 11, 12
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (October 2021) 3, 4
31. Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.Y.: Full-frame video stabilization with motion inpainting. TPAMI **28**(7), 1150–1163 (2006). https://doi.org/10.1109/TPAMI.2006.141 2, 3
32. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: ICCV. pp. 2906–2917 (October 2021) 4
33. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: ICCVW (Oct 2019) 2
34. Newson, A., Almansa, A., Fradet, M., Gousseau, Y., Pérez, P.: Video inpainting of complex scenes. Siam journal on imaging sciences **7**(4), 1993–2019 (2014) 2, 3
35. Oh, S.W., Lee, S., Lee, J.Y., Kim, S.J.: Onion-peel networks for deep video completion. In: ICCV. pp. 4403–4412 (2019) 4, 9, 10
36. Ouyang, H., Wang, T., Chen, Q.: Internal video inpainting by implicit long-range propagation. In: ICCV (2021) 4
37. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544 (2016) 2
38. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical VQ-VAE. In: CVPR. pp. 10775–10784 (2021) 2
39. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: ICCV. pp. 5533–5541 (2017) 2, 5
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015) 2, 5

41. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020) 9
42. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: CVPR. pp. 4489–4497 (2015) 5
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017) 2, 4
44. Wang, C., Huang, H., Han, X., Wang, J.: Video inpainting by jointly learning temporal structure and spatial details. In: AAAI. vol. 33, pp. 5232–5239 (2019) 4
45. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: CVPR. pp. 5463–5474 (2021) 4
46. Wang, X., Zhang, S., Qing, Z., Shao, Y., Zuo, Z., Gao, C., Sang, N.: Oadtr: Online action detection with transformers. In: ICCV. pp. 7565–7575 (October 2021) 4
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. TIP **13**(4), 600–612 (2004) 9
48. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: ICCV. pp. 22–31 (October 2021) 4
49. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 9
50. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: CVPR. pp. 3723–3732 (2019) 1, 2, 3, 9, 10, 11, 12, 13
51. Xu, S., Liu, D., Xiong, Z.: E2I: Generative inpainting from edge to image. TCSVT (2020) 2
52. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal attention for long-range interactions in vision transformers. In: NeurIPS (December 2021), `https://www.microsoft.com/en-us/research/publication/focal-self-attention-for-local-global-interactions-in-vision-transformers/` 3, 4, 9
53. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR. pp. 5505–5514 (2018) 2, 11
54. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV. pp. 4471–4480 (2019) 2
55. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: ECCV. pp. 528–543 (2020) 2, 4, 9, 10, 11, 12
56. Zhang, H., Mai, L., Xu, N., Wang, Z., Collomosse, J., Jin, H.: An internal learning approach to video inpainting. In: ICCV. pp. 2720–2729 (2019) 4
57. Zhang, K., Fu, J., Liu, D.: Inertia-guided flow completion and style fusion for video inpainting. In: CVPR. pp. 5982–5991 (June 2022) 2
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018) 9
59. Zou, X., Yang, L., Liu, D., Lee, Y.J.: Progressive temporal feature alignment network for video inpainting. In: CVPR (2021) 4, 9, 10, 11