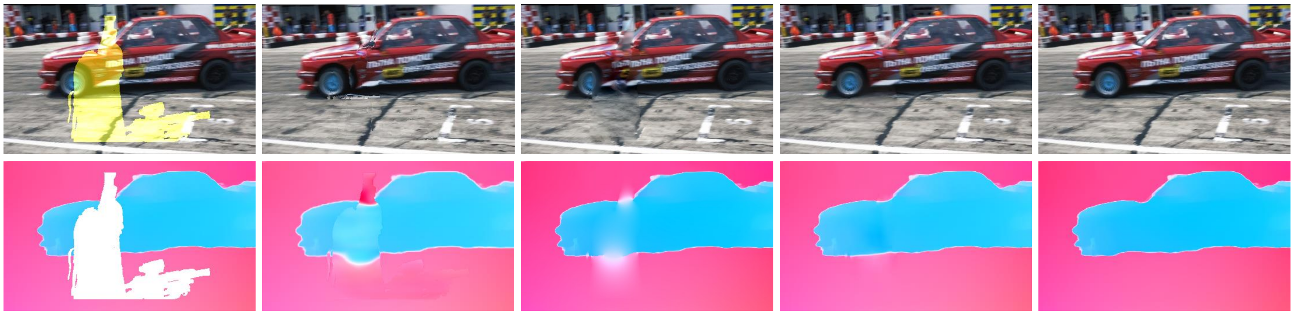
Inertia-Guided Flow Completion and Style Fusion for Video Inpainting

Kaidong Zhang[1] Jingjing Fu[2] 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) Input (b) DFGVI [48] (c) FGVC [8] (d) Ours (e) Ground-truth

Figure 1. Comparison between our results and previous flow-guided video inpainting results. Our method enjoys a visually more pleasing result with respect to the completed flow (bottom row) and the inpainted video frame (top row).

Abstract

Physical objects have inertia, which resists changes in the velocity and motion direction. Inspired by this, we introduce inertia prior that optical flow, which reflects object motion in a local temporal window, keeps unchanged in the adjacent preceding or subsequent frame. We propose a flow completion network to align and aggregate flow features from the consecutive flow sequences based on the inertia prior. The corrupted flows are completed under the supervision of customized losses on reconstruction, flow smoothness, and consistent ternary census transform. The completed flows with high fidelity give rise to significant improvement on the video inpainting quality. Nevertheless, the existing flow-guided cross-frame warping methods fail to consider the lightening and sharpness variation across video frames, which leads to spatial incoherence after warping from other frames. To alleviate such problem, we propose the Adaptive Style Fusion Network (ASFN), which utilizes the style information extracted from the valid regions to guide the gradient refinement in the warped regions. Moreover, we design a data simulation pipeline to reduce the training difficulty of ASFN. Extensive experiments show the superiority of our method against the state-of-the-art methods quantitatively and qualitatively. The project page is at https://github.com/hitachinsk/ISVI.

1. Introduction

Video inpainting aims at filling-in the corrupted regions across video frames to maintain the visual coherence of the restored video [2]. It has wide application scenarios, such as object removal, watermark removal, video retargeting, etc. Different from image inpainting [15, 31, 32, 49], video inpainting highly depends on the utilization of the complementary content across video frames to synthesize video frames with high visual quality.

Over the past two decades, researchers have committed significant efforts to video inpainting [10, 27, 36, 42, 45]. In recent years, a number of deep learning-based video inpainting methods are proposed, and they can be classified into two categories. The first category [5, 6, 12, 18, 20, 21, 29, 42, 54, 57] synthesizes the pixels in the video frames directly, while the second category [8, 48] completes the optical flows to guide the warping procedure from the valid regions to fill in the corrupted regions. We refer these two categories to pixel-based methods and flow-based methods, respectively. Compared with pixel-based methods, flow-based methods are capable of maintaining high-frequency details in the inpainted video frames, because they mainly rely on warping video frames rather than synthesizing the pixels. Therefore, flow-based methods could achieve more visual pleasing results against the pixel-based rivals [38].

This work is supported by the Natural Science Foundation of China under Grant 62036005 and Grant 62022075, and by the Fundamental Research Funds for the Central Universities under Grant WK3490000006. K. Zhang performed this work at Microsoft Research Asia.

Similar to frames, consecutive optical flows are also correlated. The fully utilization of the context provided by the flows nearby is crucial for accurate flow completion. DFGVI [48] directly concatenates the consecutive flows for target flow completion and lacks of insightful modeling on the motion correlation between the flows.

The existing flow-based methods suffer from inaccurate flow completion, which results in erroneous warping and inpainting performance degradation, observable as the seams and ghosting shown in Fig. 1b and 1c. Moreover, the style (including lightening and sharpness) across different video frames is not exactly the same, which causes spatial incoherence between the valid regions and the warped regions (the corrupted regions filled with the warped content). Although FGVC [8] has introduced gradient warping and Poisson blending [33] to obtain seamless fusion, such strategy is inadequate to deal with the style difference between each frames.

For more effective flow context utilization, we introduce the *inertia prior* for accurate flow completion in a local temporal sequence. Inertia is the resistance of any physical object to any change in its speed or direction of motion. In a local temporal window, inertia guarantees strong coherence of optical flows. Therefore, We align the features from consecutive optical flows under the inertia prior and generate richer temporal context representation, which empowers accurate flow completion. We refer this flow completion network as **I**nertia-**G**uided **F**low **C**ompletion (IGFC) Network. We also introduce the smoothness loss and the ternary census transform (TCT) loss to supervise the completion of optical flows with respect to their intrinsic properties.

To amend the spatial incoherence caused by style variation across different video frames after flow-guided warping, we design **A**daptive **S**tyle **F**usion **N**etwork (ASFN) to optimize the warped gradients in the warped regions under the guidance of the gradients in the valid regions. ASFN is a lightweight network with several adaptive style fusion (ASF) modules. In each ASF module, the mean and standard deviation of the valid regions and the warped regions are extracted and fused to correct the style in the warped regions. Experimental results demonstrate the effectiveness of ASFN in style correction for better spatial coherence.

For the training of ASFN, we design a data simulation pipeline to ease the cost on data preparation and enable separative training scheme. Besides, our method achieves memory-efficient inference and is capable to tackle videos up to 4K.

The contributions of this work can be summarized as:

- We introduce the inertia prior to model the inherent correlation within optical flow sequences, and propose the flow completion network (IGFC) with inertia-guided flow feature alignment and aggregation for high-quality flow completion.
- We propose the Adaptive Style Fusion Network (ASFN) to refine the warped gradients in the warped regions to alleviate the spatial incoherence caused by style variation across different video frames.
- We establish a data simulation pipeline for ASFN training, which degrades the data preparation cost significantly for more efficient training.

2. Related Work

Image Inpainting. Prior to the prevalence of deep learning, diffusion-based methods [3] and patch-based methods [1] are two major solutions for image inpainting. Thereafter, deep learning-based image inpainting methods emerge and they utilize powerful semantic analysis ability of CNN and GAN [9] to inpaint the corrupted images [15, 31, 32, 51]. Partial convolution [24] and gated convolution [52] are proposed to inpaint the free-form holes. Recently, researchers introduce the structure guidance [28, 49] and the semantic guidance [22] to further improve the performance of image inpainting.

Video Inpainting. Traditional methods [7, 10] complete the corrupted regions of the target frames with the valid regions from the aligned reference frames under the guidance of homography or optical flow warping. Huang *et al*. [13] propose to optimize both optical flow reconstruction and frame inpainting simultaneously to maintain the spatiotemporal consistency, which achieves excellent performance.

Recently, more methods adopt CNN in video inpainting. A number of methods [5, 6, 42, 57] adopt 3D CNN [40] or channel shift [23] for spatiotemporal joint optimization. Several studies [18, 21] introduce recurrent networks [46] to exploit the temporal relationship explicitly. Some works [30, 55] adopt the internal learning to exploit the spatiotemporal redundancy in videos, while some works adopt the attention mechanism [12, 17, 20, 25, 26, 29, 54] to fetch similar content in the feature domain for video inpainting.

Xu *et al*. [48] and Gao *et al*. [8] implement content propagation using optical flow for video inpainting. Since the videos are filled with the valid pixels under the guidance of the completed optical flows, flow-based methods are good at maintaining the spatial high-frequency details. However, the above two methods fail to explicitly consider the motion correlation between consecutive flows during optical flow completion, which leads to sub-optimal flow completion quality. Besides, style variation between different video frames causes spatial incoherence in the warped videos. Our method approximates the motion correlation between consecutive flows with the inertia prior to fuse flows more accurately, and we also design ASFN to refine the style in the warped regions.

Style Transfer. AdaIN [14] extracts and maps the mean and standard deviation from one image to the other in the

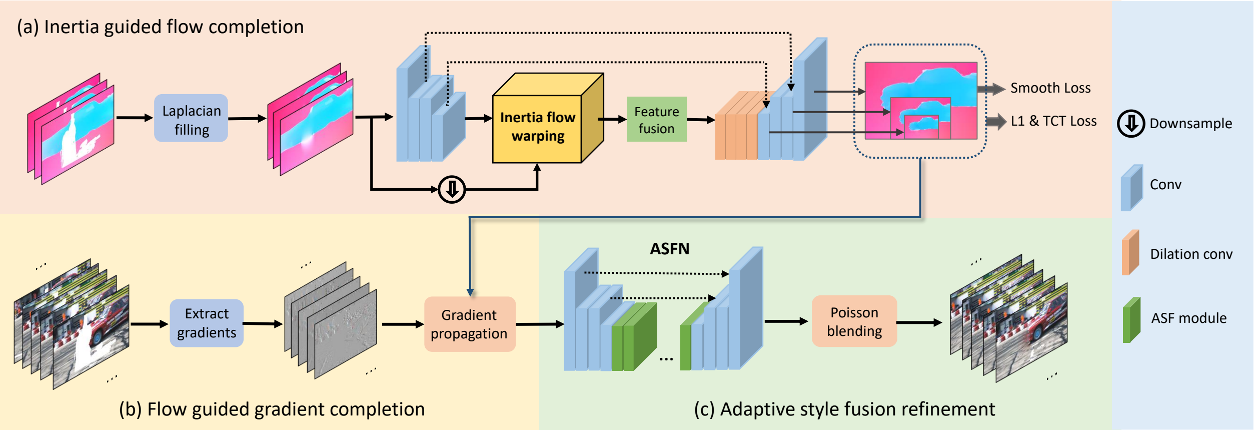


Figure 2. Our method consists of three steps. We mark these steps with different background colors. In the first step, optical flow features are aligned under the inertia prior and then fused for flow completion. Next, we utilize the completed flows with original resolution to guide the gradient propagation across the video. In the third step, we adopt ASFN to correct the style of the filled gradients based on the style in the valid regions. Finally, we use Poisson blending to render the results.

deep feature space. StyleGAN [16] controls the feature in different levels to synthesize high-quality images. Neither of the methods has been employed in video inpainting to reduce the inter-frame style difference. Our ASFN extracts feature distribution in the valid regions to guide the style refinement of the warped regions in the gradient domain.

3. Method

3.1. Problem Formulation

Given a video sequence $X := \{X_1, X_2, ..., X_t\}$, our goal is to synthesize the corrupted content indicated by the corresponding mask sequence $M := \{M_1, M_2, ..., M_t\}$, where "1" represents the corrupted regions, and "0" represents the valid regions.

3.2. Inertia-Guided Flow Completion Network

Our inertia-guided flow completion network (IGFC) is depicted in Fig. 2 (a), which is an encoder-decoder network with skip connection [35]. Here we take forward flow completion as an example to introduce flow completion in details. In this subsection, we denote the forward optical flow between t-th and $(t+1)$-th frames as F_t for simplicity. We corrupt the flow sequence $\{F_{t-i}, ..., F_t, ..., F_{t+i}\}$ with their corresponding masks, and initialize these flows with Laplacian filling, where the initialized t-th flow is denoted as $\tilde{F}_t$. The input of IGFC is the consecutive initialized flows $\{\tilde{F}_{t-i}, ..., \tilde{F}_t, ..., \tilde{F}_{t+i}\}$ and the output is the completed target flow $\hat{F}_t$. We adopt the inertia prior to align the encoded reference flow features to the target flow feature, and then fuse the features from the aligned reference features to the target feature with a matching network, as depicted by [43]. We adopt the dilation convolution [50] to enlarge the receptive field of the fused target flow. The proposed network generates optical flows in a coarse to fine manner with multiscale motion field description.

Inertia Prior assumes the motion trend in a local temporal window is constant. Given two optical flows $\tilde{F}_{t-1}$ and $\tilde{F}_t$. For a point x_{i-1} in frame I_{t-1}, inertia prior indicates,

$$\tilde{F}_t(x_{i-1} + \tilde{F}_{t-1}(x_{i-1})) \approx \tilde{F}_{t-1}(x_{i-1}) \tag{1}$$

As the flow is floating-point number, we quantify the propagated flow value to the four nearest integers based on the bilinear kernel. Given the warped pixel location $p(x_i) = x_{i-1} + \tilde{F}_{t-1}(x_{i-1})$, this process can be written as,

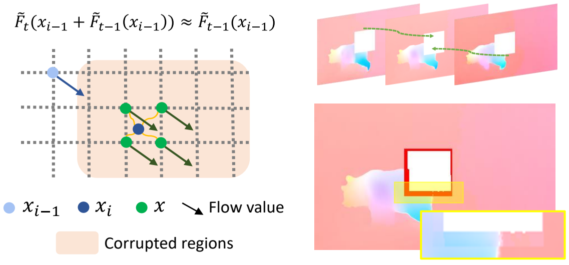
$$\tilde{F}_t(x_i) = \frac{\sum_{x \in S} k(x - p(x_i))\tilde{F}_{t-1}(x_{i-1})}{\sum_{x \in S} k(x - p(x_i))} \tag{2}$$

where $k(a) = (1 - a_x) \cdot (1 - a_y)$ is the bilinear kernel, and a_x and a_y are the coordinates of the point a. S denotes the 4-neighbor of x_i.

Inertia prior can not only align the flows nearby, but also align the flows within a certain interval. If we warp the optical flow $\tilde{F}_{t-j}$ to the flow $\tilde{F}_t$, the corresponding warped pixel from x_{t-j} is $p(x_t) = x_{t-j} + j \times \tilde{F}_{t-j}(x_{t-j})$. For the optical flows warped from the future timestamp, the corresponding j is negative. For inertia flow warping of the backward flows, we reverse the order of the the backward flows, and the above formulas still hold.

We illustrate inertia prior in Fig. 3 (a) and provide an example on pixel domain inertia warping in Fig. 3 (b). The valid regions of reference flows are transformed by inertia warping and aligned to the corrupted regions of the target flow. Such complementary features at the same location provide a good reference to the completion of the target flow. Therefore, we calculate the flow-wise similarity with a matching net to aggregate these aligned flows.

In general, the mask regions do not shift too much in a local temporal window, and hence applying inertia prior


$\tilde{F}_t(x_{i-1} + \tilde{F}_{t-1}(x_{i-1})) \approx \tilde{F}_{t-1}(x_{i-1})$
x_{i-1}
x_i
x
Flow value
Corrupted regions

(a) Inertia assumption illustration (b) Flow warping example

Figure 3. Inertia prior illustration. The red regions in (b) represents the newly filled regions propagated from the reference flows.

to align the optical flows directly cannot provide sufficient convincing references in the inner corrupted regions of the target flow. Therefore, we apply the inertia prior in the feature domain, considering that the encoding process can also be regarded as a pre-filling process. The inertia warping in the feature space can get supervised from the valid regions even in the inner corrupted regions. Moreover, inertia warping in feature space can be optimized jointly with the network, which boosts the flow completion performance.

Loss Function. IGFC outputs optical flows in a coarse to fine manner. We penalize the predicted flows at each resolution with the reconstruction loss in hole and valid regions.

$$\begin{aligned} L_{hole} &= \left\| M_t \odot (F_t - \hat{F}_t) \right\|_1 / \left\| M_t \right\|_1 \\ L_{valid} &= \left\| (1 - M_t) \odot (F_t - \hat{F}_t) \right\|_1 / \left\| (1 - M_t) \right\|_1 \end{aligned} \tag{3}$$

where $\odot$ represents Hadamard product.

Warping accuracy could supervise flow completion from the perspective of flow quality. With the completed flows, we warp the ground-truth frames after ternary census transform [37, 53], which excludes the interference caused by lightening variation across different video frames. We penalize the inaccurate warping regions with ternary census transform loss (TCT loss), denoted as L_{ter}. TCT loss is imposed to all the resolutions of the completed flows to guide multiscale motion field. Details of the TCT loss are provided in the supplementary material.

We apply the first-order and the second-order smooth losses to the completed optical flow at the original resolution for preserving its piece-wise smooth property.

$$L_{smooth} = \left\| \nabla \hat{F}_t \right\|_1 + \left\| \triangle \hat{F}_t \right\|_1 \tag{4}$$

where ∇ represents the gradient operator, and $\triangle$ represents the divergence operator.

Therefore, the loss function to train IGFC is the combination of the above four loss terms.

$$L = \lambda_1 L_{hole} + \lambda_2 L_{valid} + \lambda_3 L_{smooth} + \lambda_4 L_{ter} \tag{5}$$

We set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.5$ and $\lambda_4 = 0.01$.

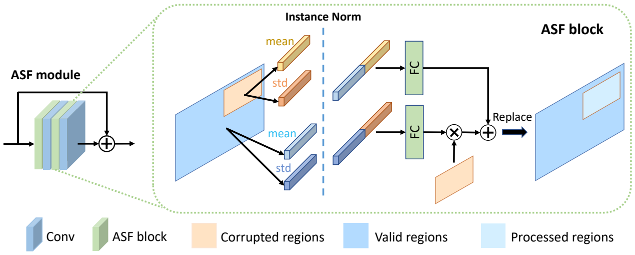


Figure 4. The structure of ASF module (left) and ASF block (right), in which the channel dimension is omitted for simplicity.

3.3. Adaptive Style Fusion Network

After we complete the optical flows, we can propagate the content across different frames along the trajectory formed by the completed optical flows. Since the completed optical flows are not perfect, the incorrect warping will misplace the reference pixels and lead to significant mismatching in low-frequency components. Therefore, we choose to propagate the gradients to avoid the low-frequency propagation error and maintain the local content consistency. The missing low-frequency components will be synthesized by Poisson blending [33] with the assistance of valid regions. Different from FGVC [8], we construct the Poisson equation based on the corrupted regions and their 2-pixel boundaries, which maintains the original performance but runs faster because of the reduction of the dimension in Poisson equation. Following previous flow-guided video inpainting methods, we inpaint the occluded regions with DeepFillV1 [51]. Our warping procedure is borrowed from FGVC [8], and more details can be viewed in the supplementary material.

Due to lightening and sharpness variation across video frames, even if the trajectory is formed by the perfect optical flows, the content propagated along the ideal trajectory cannot be guaranteed the same with the ground truth gradient in the target frame. For example, frame I_j can be written as $I_j = aC_j + b$, where C_j is the content to be propagated. a and b represent the multiplicative and additive style parameters, respectively. If we propagate gradient from $I_j(x_j)$ to $I_i(x_i)$, the gradient will be,

$$\begin{aligned} \nabla I_i(x_i) &= I_j(x_j + 1) - I_j(x_j) \\ &= aC_j(x_j + 1) + b - (aC_j(x_j) + b) \\ &= a\nabla C_j(x_j) \end{aligned} \tag{6}$$

The presence of the style parameter a in Equation 6 affects the distribution of the propagated gradient, which causes the style deviation of the gradients in the warped regions. For example, if $\nabla C_j(x_j)$ subjects to Gaussian distribution $N(\mu, \sigma^2)$, the distribution of $a\nabla C_j(x_j)$ will be $N(a\mu, a^2\sigma^2)$. The style parameter a impacts the warped gradient features of other frames through gradient propagation, which leads to the spatial incoherence.

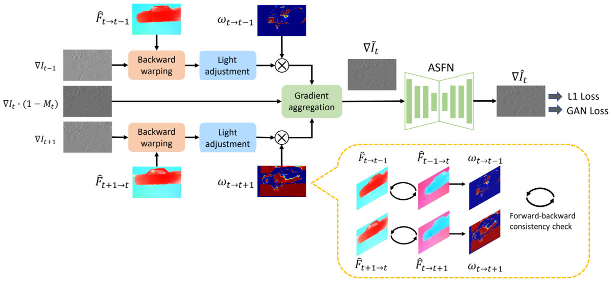


Figure 5. Data simulation pipeline for ASFN training. We complete the flows within temporal window $[t-1, t+1]$ and generate the fusion weights based on the completed flows under forward-backward consistency check. We fill the corrupted gradient of I_t with the gradients warped from nearby frames using completed flows to get the training data $\nabla\tilde{I}_t$. The warped gradients are fused with the generated fusion weights $\omega_{t\to t-1}$ and $\omega_{t\to t+1}$.

Therefore, we design ASFN to correct distribution of the warped regions. As illustrated in Fig. 2 (c), ASFN includes several ASF modules, and an ASF module consists of two ASF blocks and two convolution blocks with residual connection [11]. More ASF details are shown in Fig. 4.

We map the warped gradient map $\nabla\tilde{I}_t$ to the feature space with an encoder and process the encoded feature with four ASF modules. For the m-th ASF block, we denotes its input as p_m and the output as p_{m+1}. Given corresponding mask M_t, we extract the mean and standard deviation of p_m in the warped and the valid regions, respectively.

$$\mu_\Omega(m) = \frac{1}{\|\Omega\|}\sum_\Omega p_m$$
$$\sigma_\Omega(m) = \frac{1}{\|\Omega\|}\sqrt{\sum_\Omega (p_m - \mu_\Omega(m))^2} \tag{7}$$

where μ_Ω and σ_Ω represent the mean and standard deviation of the feature map in the corresponding regions. For the valid regions, $\Omega = 1 - M_t$, otherwise $\Omega = M_t$. $\|\Omega\|$ represents the number of pixels in Ω.

Since the style in the valid regions is known, we can adopt such style to optimize the counterpart in the warped regions. The style in the valid regions does not encode any content about the warped regions. If we directly map the style from valid regions to warped regions, the temporal style prior encoded in the warped regions may be discarded. Therefore, we optimize the style in the warped regions by concatenating the mean and standard deviation vectors in the warped and valid regions, respectively and use two FC layers to fuse the style information across these two regions to get the multiplicative and additive style parameters γ and β. We adopt the instance normalization [41] to wipe the original style information in the warped regions. Finally, The generated style information is mapped to the warped regions, and such operation can be formulated as,

$$p_{m+1} = (1-M_t)\odot p_m + M_t \odot (\gamma\frac{p_m - \mu_{M_t}(m)}{\sigma_{M_t}(m)} + \beta) \tag{8}$$

Finally, the refined gradient $\nabla\hat{I}_t$ is generated with a decoder.

Loss Function. We adopt the reconstruction loss and the adversarial loss to train ASFN, the reconstruction loss is,

$$\begin{aligned} Ls_{hole} &= \left\|M_t \odot (\nabla I_t - \nabla\hat{I}_t)\right\|_1 / \|M_t\|_1 \\ Ls_{valid} &= \left\|(1-M_t) \odot (\nabla I_t - \nabla\hat{I}_t)\right\|_1 / \|(1-M_t)\|_1 \\ Ls_{rec} &= Ls_{hole} + Ls_{valid} \end{aligned} \tag{9}$$

where ∇I_t denotes as the ground truth gradient. We adopt the SN-PatchGAN [52] to make the distribution of the refined gradients and the ground truth as close as possible and use the hinge loss for the discriminator. We denote the adversarial loss as L_{adv}, and the loss Ls is the weighted combination of the following two loss terms. We set the weight of reconstruction loss to 1, and the weight of adversarial loss to 0.01.

3.4. Data simulation pipeline

The data preparation cost for ASFN training is expensive. To obtain the training data, we need to complete each optical flow in the video with IGFC, and warp the gradients across the whole video with completed flows until there is no unfilled regions, which is unacceptable during training.

In order to reduce data preparation cost, we propose the data simulation pipeline shown in Fig. 5. We adopt the pretrained IGFC to produce the completed optical flows in a short temporal window. To guarantee the corrupted regions are filled as much as possible, we only corrupt the gradient ∇I_t with the corresponding mask M_t. The corrupted regions are filled by the propagation from the **ground truth** gradients ∇I_{t-1} and ∇I_{t+1} with the completed flows. The fusion weights $\omega_{t\to(t-1)}$ and $\omega_{t\to(t+1)}$ are calculated by the flow forward-backward consistency, and greater weights are attached to the flow-consistent area. We warp the gradients ∇I_{t-1} and ∇I_{t+1} with the completed flows $\hat{F}_{t\to(t-1)}$ and $\hat{F}_{t\to(t+1)}$, respectively. To simulate the style variation in a frame, we impose random lightening and Gaussian blur to the warped regions. Finally, we adopt the fusion weights to fuse the warped gradients and replace the corrupted regions in the gradient ∇I_t to get the training data $\nabla\tilde{I}_t$.

4. Experiments

4.1. Settings

We adopt two common datasets for evaluation: Youtube-VOS [47] and DAVIS [4]. Youtube-VOS contains 4,453

Method	Youtube-VOS			DAVIS								
				square			object			960×600		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VINet [18]	29.83	0.9548	0.0470	28.32	0.9425	0.0494	28.47	0.9222	0.0831	-	-	-
DFGVI [48]	32.05	0.9646	0.0380	29.75	0.9589	0.0371	30.28	0.9254	0.0522	29.10	0.9249	0.0564
CPN [20]	32.17	0.9630	0.0396	30.20	0.9528	0.0489	31.59	0.9332	0.0578	-	-	-
OPN [29]	32.66	0.9647	0.0386	31.15	0.9578	0.0443	32.40	0.9443	0.0413	-	-	-
3DGC [5]	30.22	0.9607	0.0410	28.19	0.9439	0.0485	31.69	0.9396	0.0535	-	-	-
STTN [54]	32.49	0.9642	0.0400	30.54	0.9540	0.0468	32.83	0.9426	0.0524	-	-	-
TSAM [57]	31.62	0.9615	0.0314	29.73	0.9505	0.0364	31.50	0.9344	0.0478	-	-	-
FFM [25]	33.73	0.9704	0.0297	31.87	0.9652	0.0340	34.19	0.9510	0.0449	-	-	-
FGVC [8]	33.94	0.9719	0.0259	32.14	0.9667	0.0298	33.91	0.9554	0.0360	34.23	0.9607	0.0345
Ours	34.79	0.9743	0.0225	33.23	0.9729	0.0247	35.16	0.9648	0.0304	35.40	0.9659	0.0303

Table 1. Quantitative results on the Youtube-VOS and DAVIS dataset. We underline the best and the second best with red and blue font. ↓ means lower is better, while ↑ means higher is better. The missing number indicates the corresponding method fails at that resolution because of the memory limitation. We adopt the resized object mask set for video inpainting at 960×600 resolution.

videos with natural scenes. We train our network with its training set and test with its test set. DAVIS contains 150 videos, whose training set has densely annotated masks. We adopt its training set as our test set to evaluate the video inpaining performance.

Following the previous work [8], we adopt PSNR, SSIM [44], and LPIPS [56] to measure the video inpainting quality, and use end-point-error (EPE) to evaluate optical flow completion quality. We compare our method with state-of-the-art baselines, including VINet [18], DFGVI [48], CPN [20], OPN [29], 3DGC [5], STTN [54], FGVC [8], TSAM [57], and FFM (a.k.a. Fuseformer) [25].

In our experiments, RAFT [39] is employed to extract optical flows. We also adopt RAFT as the flow extractor to other flow-guided video inpainting methods [8, 48] for fair comparison. We utilize three consecutive flows as inputs of IGFC to strike the balance between efficiency and performance, and the inputs to IGFC and ASFN are both resized to 256 × 256. The initial learning rate is $1e-4$ and divided by 10 after 120k iterations. Both IGFC and ASFN are trained with the Adam optimizer [19], and the whole training process takes about 3.5 days.

4.2. Quantitative Evaluation

We report the quantitative results of our method and the baselines on the Youtube-VOS and DAVIS datasets. During inference, all video frames are resized to 432×256 without specification. For Youtube-VOS, we apply square masks for inference. For DAVIS, we adopt square masks and object masks for inference. The average size of the square masks takes about $\frac{1}{16}$ of the whole frame area. The object masks are randomly shuffled from the annotations in DAVIS. We also report the video inpainting performance under the 960×600 resolution to validate the video inpainting performance under higher resolution.

The quantitative results are shown in Tab. 1. For both Youtube-VOS and DAVIS, our method outperforms the state-of-the-art baselines by a large margin. Our method enjoys superior performance not only in the restoration metric (PSNR, SSIM), but also in the perceptual metric (LPIPS). The flow results are presented in Tab. 3. Our flow completion method also significantly advances other works. The running speed of our method is also competitive with other flow-based video inpainting methods [8, 48].

4.3. Qualitative Comparisons

We perform qualitative comparisons of our method against six competitive baselines [8, 20, 25, 29, 54, 57]. The results are shown in Fig. 6. Compared with pixel-based video inpainting methods, flow-guided methods commonly generate sharper results by avoiding spectral bias [34] in CNN. Fig. 8 compares the flow completion quality of our method and previous flow-guided methods [8, 48]. Our method enjoys more accurate flow completion quality. The accurate optical flows synthesis in IGFC and the style correction in ASFN both lead to better video inpainting performance and more visually pleasing experience.

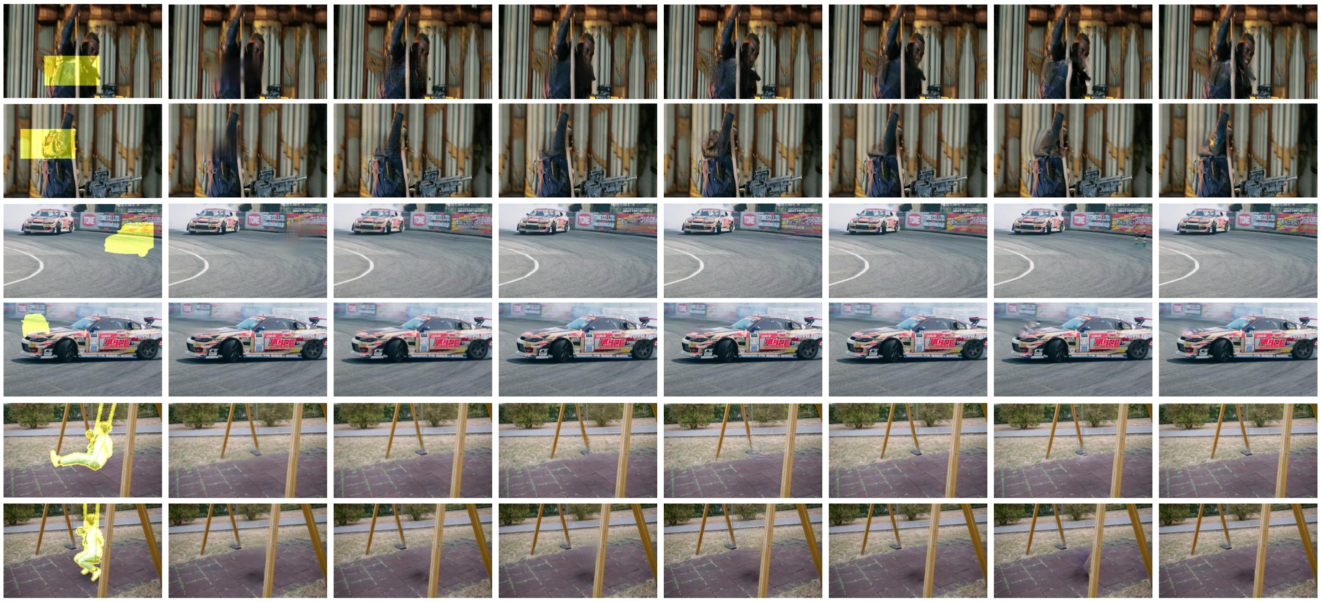
4.4. User Study

We do a user study to validate the superior subjective visual quality of our method against the others under the object removal setting. We recruit 30 volunteers. We randomly sample 20 videos from DAVIS for user study. All the videos can be replayed many times to help the volunteers make more accurate decisions. Fig. 7 shows the results between our method and the others, which illustrates the superior performance of our method.

4.5. Ablation Studies

Our ablation studies are conducted on DAVIS, and the results on both the square masks and the object masks are reported for more comprehensive evaluation.

Effectiveness of the inertia prior. We compare IGFC with two baselines. The first is our flow completion model



(a) Input (b) CPN [20] (c) OPN [29] (d) STTN [54] (e) TSAM [57] (f) FFM [25] (g) FGVC [8] (h) Ours

Figure 6. The qualitative comparison between our method and SOTAs. Compared with other results, our synthesized videos are superior in detail preserving, which leads to more visually pleasing experiences. More qualitative results can be viewed in the supplementary material.

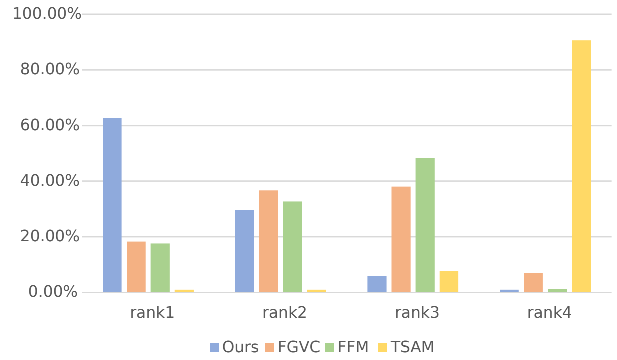
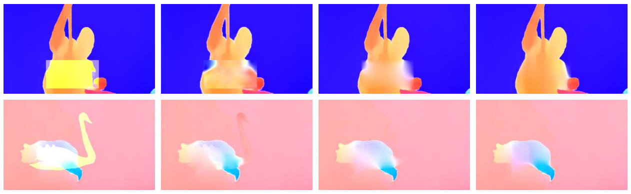


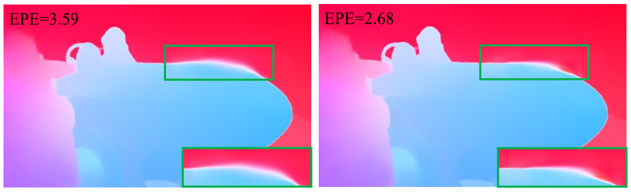
Figure 7. The user study results between our method and the competitive baselines. "Rank-x" means the percentage of the corresponding method is chosen as "x-th" best.



(a) Input (b) DFGVI [48] (c) FGVC [8] (d) Ours

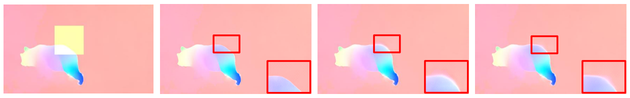
Figure 8. The comparison of the completed optical flows between IGFC and the baselines. IGFC enjoys a more accurate flow completion performance (e.g. clear motion boundary and the preservation of the details).

without inertia warping (No warp), and the second is our model with flow domain inertia warping (Flow) to validate the effectiveness of IGFC. The quantitative results are listed in Tab. 2. The feature domain inertia warping adopted by IGFC boosts both the flow completion and the video inpainting quality. The performance gain of the feature domain inertia warp mainly comes from the preservation of the motion boundary, as shown in Fig. 9. Compared with no inertia warping baseline, feature domain inertia warping provides more accurate reference to fill target flow features in the corresponding regions and accordingly benefit the flow completion. As a result, the ghost and deformation around the motion boundary get suppressed.



(a) w/o. inertia warp (b) feature domain inertia warp

Figure 9. The comparison of the completed flows w/o. inertia warping and with feature domain inertia warping. The inertia prior on feature domain can predict the motion structure and boundaries better.

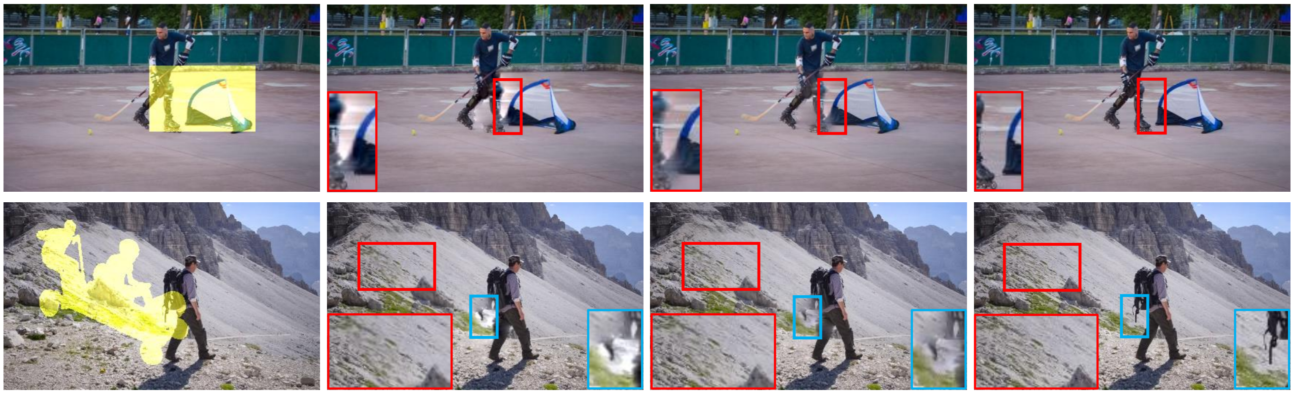


(a) Input (b) GT (c) w/o. L_{ter} (d) with L_{ter}

Figure 10. The visualization of optical flows synthesized by the models trained with or w/o. TCT loss. Compared with the flows without TCT loss supervision, the model with TCT loss can maintain the sharpness at the edges.

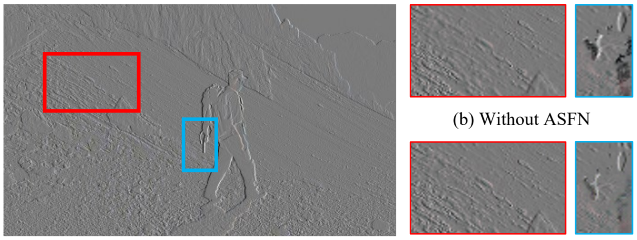
Effectiveness of the TCT loss. TCT loss supervises the flow completion quality with the frame warping after ternary census transform. Fig. 10 shows that the TCT loss is beneficial to the clarity in the motion boundary.

Effectiveness of the ASFN. Fig. 11 illustrates our in-



(a) Input (b) w/o. ASFN (c) with ASFN (d) GT

Figure 11. The comparison of the frames processed with or w/o. ASFN. ASFN can correct the unreasonable lightening around the player and the back of the hiker, and can also enhance the details in the mountain region so as to achieve spatial coherence between the warped and the valid regions.



(b) Without ASFN

(a) GT gradient (c) With ASFN

Figure 12. Comparison of the gradient map with or w/o ASFN. In the red box, ASFN sharpens the texture of the mountain based on the sharpness in the valid regions; in the blue box, ASFN reversely suppresses the over-sharp patterns for spatially coherent style.

Method	square				object			
	EPE↓	PSNR↑	SSIM↑	LPIPS↓	EPE ↓	PSNR↑	SSIM↑	LPIPS↓
No warp	0.58	32.94	0.9716	0.0267	0.39	34.90	0.9634	0.0320
Flow	0.58	32.91	0.9715	0.0269	0.38	34.96	0.9637	0.0316
IGFC	0.56	33.23	0.9729	0.0247	0.35	35.16	0.9648	0.0304

Table 2. Comparisons of the flow warping methods. “No warp” indicates the flow completion network without flow alignment, “Flow” represents the inertia prior based flow warping in the flow domain. “IGFC” indicates our proposed method.

Method	ASFN	square				object			
		EPE↓	PSNR↑	SSIM↑	LPIPS↓	EPE↓	PSNR↑	SSIM↑	LPIPS ↓
DFGVI [48]		1.16	31.24	0.9637	0.0295	1.05	33.12	0.9480	0.0392
	✓		31.22	0.9634	0.0299		33.23	0.9487	0.0386
FGVC [8]		0.63	32.14	0.9667	0.0298	0.49	33.91	0.9554	0.0360
	✓		32.37	0.9677	0.0271		34.17	0.9560	0.0351
Ours		0.56	32.91	0.9711	0.0261	0.35	34.88	0.9632	0.0322
	✓		33.23	0.9729	0.0247		35.16	0.9648	0.0304

Table 3. Comparisons of the flow completion quality and the video inpainting performance with or w/o. ASFN across different flow-guided video inpainting methods.

painting results with or without ASFN. With ASFN, our results are more spatially coherent thanks to the correction of the abnormal lightening variation (e.g. the leg of the player and the back of the hiker) and sharpness inconsistency (the texture in the mountain) in the warped regions, which are mainly caused by style variation cross different frames and inaccurate flow warping. Fig. 12 shows the gradient results of the “hike” sequence in DAVIS. We can observe that ASFN does not simply blur the gradients, but correct the style in the warped regions with the global counterparts provided by the valid regions.

Moreover, ASFN is also beneficial to other flow-guided video inpainting frameworks [8,48]. We replace IGFC with the flow completion component from the previous frameworks. The quantitative results are shown in Tab. 3. We observe that ASFN boosts the performance of all the flow-guided video inpainting methods, and the better flow completion quality leads to the higher performance gain. We believe that higher flow completion quality gives rise to more accurate warping, and hence reflects the style variation between the warped regions and valid regions more accurately. Both improvements contribute to the effective inference.

5. Conclusion

In this work, we propose a flow-guided video inpainting method. Based on the physical property of object motion, we introduce the inertia prior to exploit the correlation between consecutive optical flows for more accurate optical flow completion. We design the Adaptive Style Fusion Network to optimize the style of the warped regions under the guidance from the valid regions. Extensive experiments have demonstrated that our method performs high-quality video inpainting. In general, our method could handle flow completion of structured contents, but it still needs to ameliorate the performance on fine-grained flow completion and in fast motion cases. We improve the capabilities of video inpainting and produce more plausible results. This may have a potential negative impact that the inpainted videos may fool people with fake messages.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *TOG*, 28(3), Aug. 2009. 2

[2] Marcelo Bertalmio, Andrea L. Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *CVPR*, volume 1, pages 355–362, 2001. 1

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, page 417–424, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2

[4] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 5

[5] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3D gated convolution and temporal PatchGAN. In *ICCV*, pages 9066–9075, 2019. 1, 2, 6

[6] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Learnable gated temporal shift module for deep video inpainting. In *BMVC*, 2019. 1, 2

[7] M. Ebdelli, O. Le Meur, and C. Guillemot. Video inpainting with short-term windows: Application to object removal and error concealment. *TIP*, 24(10):3034–3047, 2015. 2

[8] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, pages 713–729, 2020. 1, 2, 4, 6, 7, 8

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, volume 27, 2014. 2

[10] Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *ECCV*, pages 682–695, 2012. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[12] Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G. Schwing. Proposal-based video completion. In *ECCV*, pages 38–54. Springer, 2020. 1, 2

[13] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *TOG*, 35(6):196:1–11, 2016. 2

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[15] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *TOG*, 36(4):107:1–14, 2017. 1, 2

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3

[17] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Occlusion-aware video object inpainting. In *ICCV*, 2021. 2

[18] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *CVPR*, pages 5792–5801, 2019. 1, 2, 6

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[20] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *ICCV*, pages 4413–4421, 2019. 1, 2, 6, 7

[21] Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *ECCV*, page 728–743, 2020. 1, 2

[22] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Image inpainting guided by coherence priors of semantics and textures. In *CVPR*, pages 6539–6548, June 2021. 2

[23] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 2

[24] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, pages 85–100, 2018. 2

[25] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, 2021. 2, 6, 7

[26] Ruixin Liu, Zhenyu Weng, Yuesheng Zhu, and Bairong Li. Temporal adaptive alignment network for deep video inpainting. In *IJCAI*, pages 927–933, 2020. 2

[27] Y. Matsushita, E. Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *PAMI*, 28(7):1150–1163, 2006. 1

[28] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure guided image inpainting using edge prediction. In *ICCVW*, Oct 2019. 2

[29] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *ICCV*, pages 4403–4412, 2019. 1, 2, 6, 7

[30] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *ICCV*, 2021. 2

[31] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 1, 2

[32] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In *CVPR*, pages 10775–10784, 2021. 1, 2

[33] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *TOG*, 22(3):313–318, July 2003. 2, 4

[34] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *ICML*, pages 5301–5310. PMLR, 2019. 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3

[36] T. Shiratori, Y. Matsushita, Xiaoou Tang, and Sing Bing Kang. Video completion by motion field transfer. In *CVPR*, volume 1, pages 411–418, 2006. 1

[37] Fridtjof Stein. Efficient computation of optical flow using the census transform. In *Joint Pattern Recognition Symposium*, pages 79–86. Springer, 2004. 4

[38] Ryan Szeto and Jason J. Corso. The devil is in the details: A diagnostic evaluation benchmark for video inpainting. *arXiv preprint arXiv:2105.05332*, 2021. 1

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 6

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *CVPR*, pages 4489–4497, 2015. 2

[41] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[42] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, volume 33, pages 5232–5239, 2019. 1, 2

[43] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong Li. Displacement-invariant matching cost learning for accurate optical flow estimation. In *NIPS*, volume 33, 2020. 3

[44] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 6

[45] Y. Wexler, E. Shechtman, and M. Irani. Space-time video completion. In *CVPR*, volume 1, pages I–I, 2004. 1

[46] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, pages 802–810, 2015. 2

[47] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5

[48] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *CVPR*, pages 3723–3732, 2019. 1, 2, 6, 7, 8

[49] Shunxin Xu, Dong Liu, and Zhiwei Xiong. E2I: Generative inpainting from edge to image. *TCSVT*, 2020. 1, 2

[50] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

[51] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018. 2, 4

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 2, 5

[53] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, pages 151–158. Springer, 1994. 4

[54] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, pages 528–543, 2020. 1, 2, 6, 7

[55] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *ICCV*, pages 2720–2729, 2019. 2

[56] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[57] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *CVPR*, 2021. 1, 2, 6, 7