Structure-Guided Deep Video Inpainting

Chaoqun Wang, Xuejin Chen, *Member, IEEE*, Shaobo Min, Jiaping Wang, *Member, IEEE*, and Zheng-Jun Zha *Member, IEEE*

***Abstract*—A fundamental challenge in video inpainting is the difficulty of generating video contents with fine details, while keeping spatio-temporal coherence in the missing region. Recent studies focus on synthesizing temporally smooth pixels by exploiting the flow information, while ignoring maintaining the semantic structural coherence between frames. This makes them suffer from over-smoothing and blurry contours, which significantly reduces the visual quality of inpainting results. To address this issue, we present a novel structure-guided video inpainting approach that enhances temporal structure coherence to improve video inpainting results. In contrast to directly synthesizing the missing pixel colors, we first complete the edges in the missing regions to depict scene structures and object shapes via an edge inpainting network with 3D convolutions. Then, we replenish textures using a coarse-to-fine synthesis network with a structure attention module (SAM), under the guidance of the completed edges. Specifically, our SAM is designed to model the semantic correlation between video textures and structural edges to generate more realistic contents. Besides, flow between neighboring frames is employed to enhance temporal consistency for self-supervision during training the edge inpainting and texture inpainting modules. Consequently, the inpainting results using our approach are visually pleasing with fine details and temporal coherence in low computational cost. Experiments on the YouTubeVOS and DAVIS datasets show that our method obtains state-of-the-art performance under multiple different video inpainting settings.**

***Index Terms*—Video inpainting, Structure guidance, Flow Assistance.**

I. Introduction

Video inpainting aims to recover the missing content of a corrupted video and assists lots of practical applications, *e.g.,* video restoration and watermarking removal. High-quality video inpainting requires not only realistic structures with visual details but also temporal consistency. Though great progress has been made in 2D image inpainting using deep learning techniques [1], [2], [3], directly applying these approaches to each frame individually for video inpainting will lead to flaws, flickers, and jitters due to the additional time dimension.

Traditional video inpainting methods employ a patch composition framework that composites visually pleasing content in the missing regions via patches by exploiting complementary information across neighboring frames [4], [5], [6], [7]. These methods rely heavily on the hypothesis that the missing content in the corrupted region appears in neighboring frames, which greatly limits their generalization ability. Recently, deep-learning-based methods achieve great performance improvement in video inpainting [8], [9], [10], [11]. A straightforward solution is to utilize 3D convolution layers to extract spatio-temporal features and predict missing contents with smooth motion [8]. To obtain temporally smooth results, contextual information from neighboring frames is aggregated to synthesize corrupted regions using a recurrent feedback scheme [9], [11], or pixel propagation guided by completed flows [10]. By integrating motion guidance, these methods pay more attention to temporal smoothness; however, structure rationality and object details have not been well recovered.

Without definite representation and generation of the target image structures, these methods tend to produce over-smoothed regions. Similar observations have been obtained in the image inpainting task [2], [12]. To solve this problem, two-step methods are proposed to complete object contours [2] or edge maps [12] first as auxiliary information to guide texture synthesis later in image inpainting. However, when applying these edge-first image inpainting methods to video inpainting, it brings another challenge in generating temporally coherent structures while human vision is significantly sensitive to temporal discontinuity that frequently occurs at edges.

In order to simultaneously hallucinate detailed image structures and preserve temporal coherence in video inpainting, we present a novel structure-guided video inpainting approach which effectively exploits the spatio-temporal structure information to improve the quality of video inpainting. Compared with previous video inpainting methods that only consider motion guidance, we explore the correlation between structure, texture, and motion to complete the missing region with reasonable structure, rich visual details, and temporal coherence, as shown in Fig. 1. First, we design an edge inpainting network (ENet) to predict sparse edges in the missing region to represent the target structure for each frame by exploiting the spatio-temporal neighboring information from adjacent frames. Then, under the guidance of completed edges, we employ a texture inpainting network (TexNet) to fill the missing region via a coarse-to-fine architecture and a structure attention module (SAM). Specifically, the SAM is designed to guide the texture generation by capturing the latent spatial relevance between video textures and the completed structural edges. Notably, such a structure-texture relevance can effectively improve the inpainting quality in TexNet with fewer cracks and more realistic object contours. Furthermore, to enhance the temporal coherence of synthesized frames, we employ motion flows for consistency check of both edge maps and inpainted frames during the training stage. The ground truth optical flow is exploited to guide both ENet and TexNet to generate temporal smooth edge maps and texture results via edge consistency and frame warping losses.Consequently, the inpainted frames using our approach are not only temporally consistent, but also more complete in structure and rich in

C. Wang, X. Chen, S. Min, and Z. Zha are with the National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, University of Science and Technology of China, Hefei, 230026, China.

J. Wang is with the Peng Cheng Laboratory, Shenzhen, 518000, China.

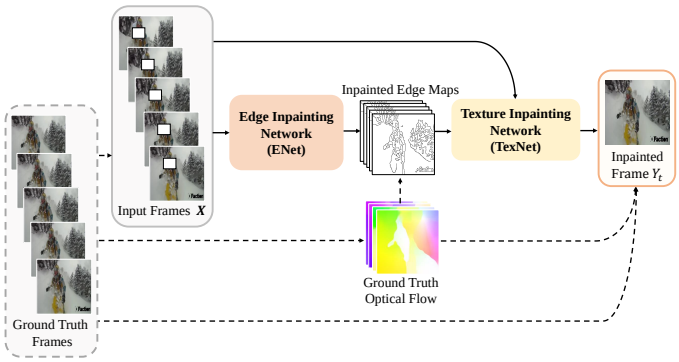

Fig. 1. Overview of our structure-guided video inpainting network. We first complete the missing edges by aggregating information from neighboring frames to represent the target structure using the ENet. Then, the TexNet synthesizes missing textures under the guidance of structure edges. Besides, the ground truth optical flow between frames is utilized during the training stage for both ENet and TexNet to enforce temporal coherence of completed contents, as illustrated by dotted lines.

visual details.

We conduct a series of experiments on the YouTubeVOS and DAVIS datasets under different mask settings. The results show that the proposed method obtains new state-of-the-art inpainting performance both quantitatively and qualitatively. Because of the sparsity of edge maps, the computational cost is also greatly reduced. In summary, our technical contributions are three-fold:

- We propose a novel structure-guided video inpainting method which integrates scene structure, texture, and motion to complete the missing region with realistic structure, rich visual details, and temporal coherence.
- A structure attention module is designed to capture the correlation between hallucinated edges and video textures, which can provide better structural guidance for texture synthesis.
- Flow-guided edge and frame consistency constraints are developed to enhance the temporal coherence of the completed edges and video frames.

II. Related Work

a) Traditional Image/Video Inpainting: Image or video inpainting has been studied for decades. Traditional methods of image and video inpainting can be divided into two categories, diffusion-based and patch-based methods. Diffusion-based methods [13], [14], [15] gradually propagate contents from surrounding areas to the missing region. However, this kind of method fails to handle large holes due to its assumption of local smoothness. Patch-based image inpainting methods, also called exemplar-based methods [16], [17], are more widely studied. They formulate the completion task as a patch-based optimization problem. Barnes *et al.* [18] employ approximate nearest neighbor algorithm to fill the damaged regions. Sangeetha *et al.* [19] propose to propagate both linear structure and two-dimensional texture into the target region. Ružić *et al.* [20] introduce Markov random field to search the most matched candidates. Ding *et al.* [21] employ nonlocal texture similarity and local intensity smoothness to produce natural-looking results. Besides, some patch-based methods utilize low rank approximation. For example, Guo *et al.* [22] propose a simple two-stage low rank approximation to recover the corrupted region, which avoids time-consuming iterations. Lu *et al.* [23] adopt gradient-based low rank approximation. These patch-based methods fill the missing content by borrowing and aggregating the most similar patches based on low-level image features from known regions. However, they usually fail when there is insufficient information in known regions or image textures are too complicated.

b) Patch-Based Video Inpainting: Patch-based methods are also widely studied for video inpainting. They search similar patches and borrow appearances from known regions across neighboring frames to synthesize the unknown content. Wexler *et al.* [24] constrain masked regions to synthesize coherent structures with respect to reference examples based on local structures. Umeda *et al.* [25] propose using directional median filter as complementation of patch-based filling. Newson *et al.* [7] extend the 2D PatchMatch algorithm [18] into 3D version to improve video inpainting quality. Huang *et al.* [26] jointly estimate optical flow and textures to promote temporal coherence. Another group of methods separate foreground and background apart, and then deal with the two parts respectively with different algorithms, since there naturally exists property differences between them. Ghanbari *et al.* [27] first separate the two parts in videos, and fills the two parts accordingly with the help of contours. Xia *et al.* [28] make use of Gaussian mixture models to also distinguish moving foreground and still background, and process them separately. However, in these patch-based video inpainting methods, the patch-searching process suffers from high computational complexity, thus limiting their usage in practical applications.

c) CNN-based Image Inpainting: Recently, deep learning methods have achieved tremendous progress in the field of computer vision. The tasks of image and video inpainting also have witnessed great promotion thanks to the capability of deep neural networks to capture high-level semantic information in images and videos. A convolution neural network (CNN) is first introduced to directly synthesize image contents in the masked regions for image denoising and inpainting in [29]. To improve the photorealism of the synthesiszed content, a generative adversarial network is employed [30]. Then, Yang *et al.* [31] take advantages of multi-scale representation to boost details generation. Multiple discriminators are used to constrain both global and local coherence of image contents [32]. Yu *et al.* [33] propose a contextual attention module to capture long-range information. Subsequent approaches solve more specific problems in image inpainting, for example, inpainting irregular holes with partial convolution [34] or employing gated convolution [1] for dynamic feature selection. While these methods tend to generate over-smoothed and blurry results, a two-stage approach is proposed to hallucinate edges first and then fill image colors using the edges as a prior [12]. Similarly, Xiong *et al.* [2] predict contours of foreground objects to guide the inpainting process of masked regions. Though high-quality static images with reasonable structure can be generated using these edge-based methods, simply

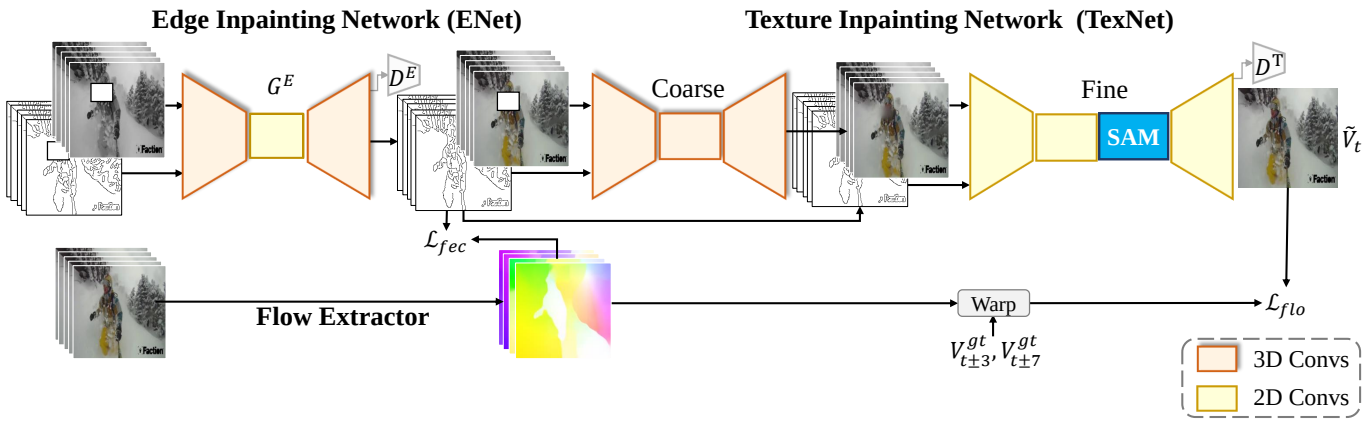

Fig. 2. The detailed architecture of our network. ENet adopts an encoder-decoder architecture to complete edges in the missing regions. TexNet utilizes a coarse-to-fine manner to inpaint the final frames. Ground truth optical flow between adjacent frames is extracted for edge consistency loss and frame warping loss.

extending it from image inpainting to video inpainting by 3D convolutions inevitably fails because of no guarantee on the temporal coherence, especially on the high-frequency signals. Besides, these methods simply utilize edges and contours as one additional channel of the image inpainting network, without exploiting a more effective mechanism to utilize the structure information more efficiently. In comparison, we design a structure attention module to better explore the structural guidance in synthesizing textures.

d) Deep Video Inpainting: Several video inpainting methods based on deep neural networks have been proposed recently, to fully utilize useful complementary information in neighboring frames. The first deep-learning-based video inpainting method is CombCN [8], which jointly learns temporal structure and spatial details via 3D convolutions. To enforce temporal coherence, features from neighboring frames are collected and refined to synthesize the missing content with recurrent feedback [9], [11]. Instead of filling pixel colors directly using CNNs, a deep flow-guided inpainting network is proposed to estimate optical flow first in the missing region and then propagate pixel colors based on the completed flow [10]. However, these existing methods usually suffer from blurs and structural cracks in the synthesized frames since it is non-trivial to maintain fine details and sharp edges at the same time with predicting temporally coherent pixel colors. In comparison, we propose to explicitly complete the target structure using edges, which are efficient to predict due to their sparsity. To utilize structural information more effectively, we also introduce a structure attention mechanism. Under the structural guidance, more visually pleasing results could be synthesized with plausible structure and fine details.

III. Approach

The target of our method is to recover the missing contents in a corrupted video with fine details and temporal consistency. We complete each frame by aggregating information for its neighboring frames. In order to synthesize the missing content and produce a completed frame $\widetilde{V}_t$ at time t, we take total T input frames $\boldsymbol{V}$ ($T = 5$), indexed by $\{V_{t-7}, V_{t-3}, V_t, V_{t+3}, V_{t+7}\}$, as input to our inpainting network in each data batch. The corrupted regions are indicated by binary masks $\boldsymbol{M}$ where $M_t^p = 1$ indicates corrupted pixel p in frame V_t.

The detailed network architecture is shown in Fig. 2. It consists of three main components. The first part is an edge inpainting network (ENet) for structure inference that recovers edge maps in the missing regions of the input frames. The second part is a coarse-to-fine texture inpainting network (TexNet) that aims to complete the missing content with visual details under the guidance of hallucinated edges. The third part leverages the ground truth motion flow as auxiliary constraints to enforce temporal coherence of both the completed edge maps and textures at the training stage.

A. Edge Inpainting Network

The edge inpainting network (ENet) completes image edges in the missing regions to depict scene structures and object shapes. Given the input corrupted frames $\boldsymbol{V}$ as well as the corresponding binary masks $\boldsymbol{M}$, a Canny edge detector is first applied to extract the corresponding edge maps $\boldsymbol{E}^i$ in the uncorrupted regions. The input of ENet consists of the incomplete grayscale version of frames $\boldsymbol{V}^g$, initial edge maps $\boldsymbol{E}^i$, and their corresponding masks $\boldsymbol{M}$. As shown in Fig. 2, our ENet, denoted by G^E, is composed of a two-layer 3D encoder, eight 2D residual blocks, and a two-layer 3D decoder. The 3D encoder and decoder are designed to learn the spatio-temporal correlation with 3D convolution operations. The intermediate 2D residual blocks are used to enlarge the spatial receptive fields by using 2D convolutions with large kernel size. Such a hybrid 3D and 2D convolution architecture can achieve a good trade-off between spatial and temporal coherence. The inpainted T edge maps $\widetilde{\boldsymbol{E}} = \{\widetilde{E}_{t-7}, \widetilde{E}_{t-3}, \widetilde{E}_t, \widetilde{E}_{t+3}, \widetilde{E}_{t+7}\}$ are obtained by:

$$\widetilde{\boldsymbol{E}} = G^E(\boldsymbol{E}^i, \boldsymbol{V}^g, \boldsymbol{M}). \tag{1}$$

To train the edge generator G^E, ENet plays a minimax game by:

$$\min_{G^E} \max_{D^E} \left(\mathcal{L}_{adv}^E + \lambda_1 \mathcal{L}_{fm}^E\right), \tag{2}$$

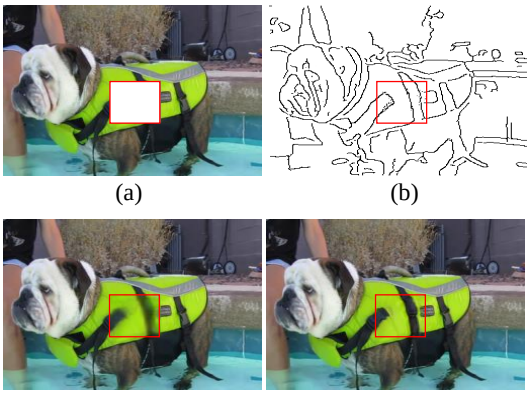

Fig. 3. To inpaint missing regions in a corrupted frame (a), our ENet first completes corresponding sparse edges (b), which well represent the structure of the missing contents. Then TexNet progressively replenishes textures under the guidance of synthesized edges from coarse (c) to fine (d).

where the discriminator D^E follows the 70×70 PatchGAN architecture [35]. $\mathcal{L}^E_{adv}$ and $\mathcal{L}^E_{fm}$ are the adversarial loss and feature matching loss respectively. λ_1 is a hyper-parameter to balance the two terms. In Eq. (2), $\mathcal{L}^E_{adv}$ is an adversarial learning loss to make the predicted edge maps more realistic, which evaluates the image-level similarity between ground truth edge maps and the predicted edge maps by:

$$\begin{aligned}\mathcal{L}^E_{adv} =& \mathbb{E}_{(E^{gt},V^g)}\big[logD^E(E^{gt},V^g)\big] \\ &+\mathbb{E}_{(\widetilde{E},V^g)}\big[log\big(1-D^E(\widetilde{E},V^g)\big)\big].\end{aligned} \tag{3}$$

$\mathcal{L}^E_{fm}$ evaluates the feature-level similarity between ground truth and predicted edge maps, which is defined by:

$$\mathcal{L}^E_{fm} = \sum_{t=1}^{T}\sum_{k=1}^{L}\frac{1}{N_k}\left\|D^E_k(E^{gt}_t,V^g_t) - D^E_k(\widetilde{E}_t,V^g_t)\right\|_1, \tag{4}$$

where D^E_k is the k-th layer in the L-layer D^E, while N_k is the element number of the output of D^E_k. By considering both feature-level and image-level similarities in Eq. (2), the edge generator G^E can be trained to produce plausible and structurally rational edge maps.

As a result, the proposed hybrid 3D and 2D convolution architecture and the two-level loss function enable the ENet to hallucinate the missing edges accurately. An example of the generated edge map is given in Fig. 3 (a) and (b). The stripe pattern on the dog body is well inferred from neighboring frames.

B. Edge-Guided Texture Inpainting Network

With the completed edge maps $\widetilde{\boldsymbol{E}}$ for the T frames $\boldsymbol{V}$, we then fill the image texture using a coarse-to-fine network‘ TexNet. Notably, the image structure, *e.g.,* object shapes, is well represented by the completed edge maps $\widetilde{\boldsymbol{E}}$. Thus, it becomes easier to fill the missing texture with the structural guidance of $\widetilde{\boldsymbol{E}}$.

To synthesize realistic frame textures, our TexNet adopts a coarse-to-fine architecture, as shown in Fig. 2. Specifically,

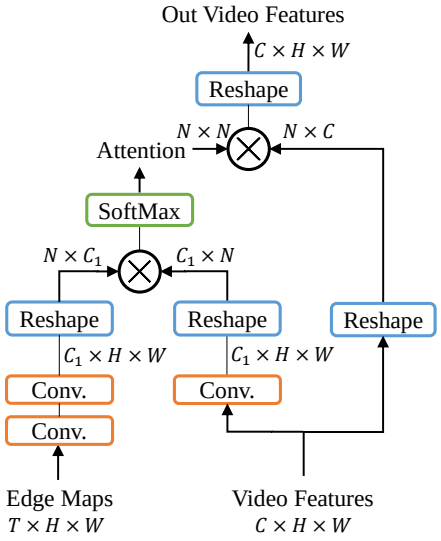

Fig. 4. Architecture of the structure attention module. C is the channel of the input video features, and $N = H \times W$. $\otimes$ represents matrix multiplication. Usually, we set $C_1 = C/8$.

TexNet consists of a coarse inpainting network and a refinement network. First, the coarse inpainting network consists of a set of 3D convolutions to capture the spatio-temporal information from the hallucinated edge maps and the corrupted input frames, and produces a rough completion $\widetilde{\boldsymbol{V}}^i$ for the T frames with colors and textures. Then, the refinement network takes the rough inpainting results $\widetilde{\boldsymbol{V}}^i$ and the synthesized edge maps $\widetilde{\boldsymbol{E}}$ as inputs to further refine texture details in $\widetilde{\boldsymbol{V}}^i$ with the guidance of structural edges $\widetilde{\boldsymbol{E}}$. Notably, only 2D convolutional layers are used in the refinement network to improve the inference efficiency.

Besides taking $\widetilde{\boldsymbol{E}}$ as an auxiliary input in the refinement network, we further design a structure attention module (SAM) to fully encode the structural information. The detailed implementation of SAM is given in Fig. 4. The inputs of SAM are the intermediate video features extracted from $\widetilde{\boldsymbol{V}}^i$ and $\widetilde{\boldsymbol{E}}$ via 2D convolutional layers, as well as the edge maps. First, the intermediate video features and embedded edge features are interacted to calculate the latent structure-texture correlation via matrix multiplication. After a SoftMax operation, the normalized attention map is obtained while conveys the correlation between sparse structures and video features. Then, the normalized attention map is applied to the intermediate video features, and the structure information is thus embedded in TexNet, which can better extract useful structural information brought by edges. After introducing structural guidance in the refinement network, the inpainted content by TexNet becomes more realistic, as Fig. 3 (d) shows.

The coarse inpainting network and the refinement network in TexNet are trained as a whole, define by G^T, end-to-end in adversarial manner by:

$$\min_{G^T}\max_{D^T}\big(\mathcal{L}^T_{rec} + \mathcal{L}^T_{adv}\big), \tag{5}$$

where D^T is the discriminator. Inspired by [12], the first term $\mathcal{L}^T_{rec}$ is the l_1-reconstruction loss to measure the difference

between predicted video frames and the ground truth video frames $\boldsymbol{V}^{gt}$. Differently, we penalize both the coarse predictions $\widetilde{\boldsymbol{V}}^{i}$ and refined frame $\widetilde{V}_t$, given by:

$$\begin{aligned}\mathcal{L}_{rec}^{T} = &\frac{1}{\|M_t\|_1}\left\|(\widetilde{V}_t - V_t^{gt})\odot M_t\right\|_1, \\ &+\lambda_2 * \frac{1}{\|\boldsymbol{M}\|_1}\left\|(\widetilde{\boldsymbol{V}}^{i} - \boldsymbol{V}^{gt})\odot \boldsymbol{M}\right\|_1,\end{aligned} \tag{6}$$

where V_t^{gt} denotes the ground truth frame at time t, and M_t is the corresponding binary mask. Besides, an extra adversarial loss $\mathcal{L}_{adv}^{T}$ is introduced in Eq. (5) to promote the visual realism of the generated frame by:

$$\mathcal{L}_{adv}^{T} = \mathbb{E}[logD^{T}(V_t^{gt})] + \mathbb{E}[log\big(1 - D^{T}(\widetilde{V}_t)\big)]. \tag{7}$$

$\mathcal{L}_{adv}^{T}$ enforces the generated frame to be more realistic.

The coarse and refinement networks can be effectively trained jointly with the combined loss of $\mathcal{L}_{rec}^{T}$ and $\mathcal{L}_{adv}^{T}$. The results in Fig. 3 (c) and (d) show that the coarse network completes the missing content with rough structures and the refinement network exactly refines the inpainting results with more sharp contours and realistic textures.

C. Flow-Guided Temporal Coherence Enhancement

Besides the structural guidance, motion information is also considered to enhance temporal consistency among the recovered frames. To this end, we employ optical flow in the training stage of both ENet and TexNet. A set of flow maps $\boldsymbol{O}$ between the current frame V_t^{gt} and its neighboring frames are generated using a pre-trained flow extraction network, such as FlowNet2.0 [36]. Specifically, $\boldsymbol{O}$ consists of four flow maps $(O_{t\Rightarrow t-7}, O_{t\Rightarrow t-3}, O_{t\Rightarrow t+3}, O_{t\Rightarrow t+7})$.

In terms of the ENet which completes the missing edges, $\boldsymbol{O}$ is used to first warp the neighboring edge maps to the current frame, and then compute the consistency between neighboring edge maps. Thus, a flow-guided edge consistency loss is defined as:

$$\mathcal{L}_{fec}^{E} = \sum_{k}\frac{1}{\|M_t\|_1}\left\|\big(\widetilde{E}_t - \phi(O_{t\Rightarrow t+k}, E_{t+k}^{gt})\big)\odot M_t\right\|_1, \tag{8}$$

where $\phi(O_{t\Rightarrow t+k}, E_{t+k}^{gt})$ is the warping operation which warps the edge map E_{t+k}^{gt} to the target frame according to the generated optical flow $O_{t\Rightarrow t+k}$. k denotes the index of neighboring frames ($k \in \{-7, -3, +3, +7\}$). With the flow-guided edge consistency loss $\mathcal{L}_{fec}^{E}$, the loss function of Eq. (2) for ENet becomes:

$$\min_{G^E}\max_{D^E}\big(\mathcal{L}_{adv}^{E} + \lambda_1 * \mathcal{L}_{fm}^{E} + \mathcal{L}_{fec}^{E}\big). \tag{9}$$

About the TexNet, we further enforce the temporal coherence of synthesized neighboring textures via a flow warping constraint $\mathcal{L}_{flo}^{T}$ by:

$$\mathcal{L}_{flo}^{T} = \sum_{k}\frac{1}{\|M_t\|_1}\left\|\big(\widetilde{V}_t - \phi(O_{t\Rightarrow t+k}, V_{t+k}^{gt})\big)\odot M_t\right\|_1, \tag{10}$$

where k is also in $\{-7, -3, 3, 7\}$. $\phi(O_{t\Rightarrow t+k}, V_{t+k}^{gt})$ warps V_{t+k}^{gt} to the target frame using flow $O_{t\Rightarrow t+k}$. Finally, $\mathcal{L}_{flo}^{T}$ is added to Eq. (5), which becomes:

$$\min_{G^T}\max_{D^T}\big(\mathcal{L}_{rec}^{T} + \mathcal{L}_{adv}^{T} + \mathcal{L}_{flo}^{T}\big). \tag{11}$$

After adding the motion information to both the training process of ENet and TexNet, the temporal consistency can be enhanced in both edge generation and texture inpainting. Since the motion is only used in the training stage, the inference of the proposed structure-guided inpainting method is efficient and effective.

IV. Experiments

To evaluate the effectiveness of our approach, we conduct a series of comparison experiments and ablation studies on two widely used datasets, *i.e.,* YouTubeVOS [37] and DAVIS [38], under different settings.

A. Experimental Settings

Mask Setting. Considering different real-world applications, we test four kinds of mask settings in this paper, which are different in shapes and positions of the missing regions.

1) Fixed square mask: The size and position of the missing square regions are fixed through the whole video.
2) Moving square mask: The position and size of the square masks change over frames.
3) Free-from mask: We apply irregular masks which imitate hand-drawn masks on each frame, following [34].
4) Foreground object mask: This type of mask is defined to line out foreground objects in videos and used for testing object removal.

Dataset. YouTubeVOS and DAVIS are widely used for evaluating video inpainting results in recent studies. YouTubeVOS consists of 4,453 video clips that contain more than 70 categories of common objects. The videos are split into three parts, 3,471 for training, 474 for validation, and 508 for testing. Since YoutubeVOS has no dense foreground mask annotations, we only use it for evaluation under mask settings (1), (2), and (3). DAVIS dataset contains 90 video sequences that are annotated with foreground object masks for test and 60 unlabeled videos for training.

Implementation details and Evaluation Metrics. All the models are tested on a TITAN X (Pascal) GPU with frame size 256×256. For training data, we randomly sample a training clip every 40 frames from each video in the dataset. Our training process consists of three steps. First, we train ENet with learning rate set as $1e-4$ for G^E and $1e-5$ for D^E. Then we train TexNet while fixing ENet. The learning rate is set as $1e-4$ for G^T, and $4e-4$ for D^T. We first train ENet and TexNet without flow-constrained losses, and then add the flow losses to finetune the networks. An Adam optimizer with $\beta = (0.9, 0.999)$ is used for all sub-network training. As for the hyper-parameters, we set $\lambda_1 = 10.0, \lambda_2 = 0.2$.

Different data preparations and evaluation metrics are used according to mask settings. We randomly generate masks for training videos in terms of mask settings (1), (2), and

Fig. 5. Comparison with state-of-the-art methods on the YouTubeVOS dataset under different mask settings. Our method produces results with more complete object structures and finer details.

(3). Masked videos are used for testing. Three commonly-used metrics, including structural similarity index (SSIM) [39], peak signal-to-noise ratio (PSNR), and Fréchet Inception Distance (FID) [40] are used to quantitatively evaluate the performance of our method. For the mask setting (4), for each training video, we randomly select a test video from the 90 test videos in the DAVIS dataset and apply its masks for the current training video with random rotation, scaling, and translation. Our inpainting network is first trained on the YoutubeVOS dataset and then finetuned on the DAVIS dataset. Since there are no ground truth videos available for this mask setting, we can not conduct quantitative evaluations to measure the output quality. In contrast, a user study is conducted for video foreground object removal.

B. Evaluation of Video Inpainting on YouTubeVOS

We compare the proposed method with four state-of-the-art video inpainting methods [12], [8], [9], [10] for the first three mask settings on the YouTubeVOS dataset. We train [12], [10] using their published codes and re-implement [8] according to their paper. As for [9], we use the officially provided model.

The quantitative results and inference speeds are reported in Table I. It shows that our method outperforms state-of-the-art methods on the three metrics, demonstrating the effectiveness of introducing structure guidance into video inpainting. Moreover, our method is also very efficient, e.g., four times faster than DVI [9] and nine times faster than DFVI [10]. Some inpainting examples are shown in Fig. 5. Compared with existing methods, the inpainting results predicted by our method are more realistic with finer details. We can observe that the frames completed using our method contain sharper object boundaries. This is achieved by the effectiveness of structure information in video inpainting. It can also be seen that our method produces temporally smooth results when observing neighboring frames.

Compared with the 2D image inpainting method Edge-Connect[12], which also predicts edges to represent the target structure, our method greatly increases the completion performance by leveraging neighboring frames to complete

TABLE I
COMPARISONS WITH FOUR STATE-OF-THE-ART METHODS ON YOUTUBEVOS. OUR METHOD OUTPERFORMS ALL OTHER METHODS ON THREE METRICS, WITH FAST INFERENCE SPEED.

	Fixed Square Mask			Moving Square Mask			Free-Form Mask			Inference
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID	Speed (fps)
Edge-Connect [12]	28.6446	0.9484	38.2116	30.7478	0.9647	16.2739	25.6693	0.9088	43.0366	22.81
CombCN [8]	27.9668	0.9515	40.7199	31.5776	0.9678	13.8383	32.1862	0.9626	19.1191	8.1634
DVI [9]	28.0846	0.9468	39.9377	36.8598	0.9728	7.2315	33.5549	0.9646	9.3797	1.2275
DFVI [10]	29.0531	0.9497	32.8860	37.8241	0.9772	6.3746	32.6287	0.9618	11.1501	0.5620
Ours	**30.0590**	**0.9543**	**27.2431**	**38.8186**	**0.9824**	**2.3455**	**35.9613**	**0.9721**	**5.8694**	5.1546

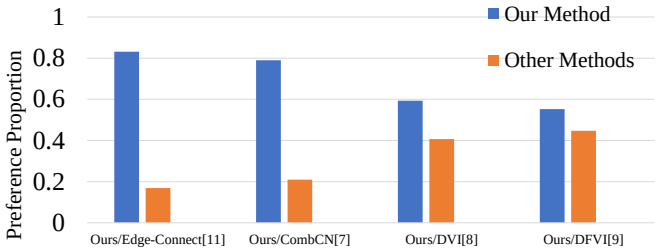

Fig. 6. Results of user study. The restored videos using our method are preferred by more participants compared with other state-of-the-art methods.

edges and synthesize textures. Thus, our method can generate more temporally coherent and realistic contents. The inference speed of Edge-Connect is faster than our method since it does not consider axillary temporal information between frames. Compared with the second-best video inpainting method DFVI [10], our method produces frames with finer structural details. Besides, only ENet and TexNet are used to directly predict final outputs in the testing period in our method, while DFVI requires iterative pixel propagation. Thus the inference speed of our method is much faster than that of DFVI. Both the quantitative and qualitative results demonstrate that our method is not a naive extension to utilize structure information in video inpainting. The well designed network architecture fully exploits structural guidance in video inpainting and brings large performance gain.

C. Object Removal on DAVIS

In regard to the foreground object mask setting that aims to remove undesired objects in videos, there is no ground truth for quantitative evaluation. Therefore, we conduct a user study on the DAVIS dataset to evaluate the visual quality of our method, compared with the four methods [12], [8], [9], [10]. In each test, we show three videos to the subject at the same time. The original video with red masks indicating objects to remove is shown in the middle, while the inpainting results of our method and one of the other four methods are shown on the two sides in random order. The subjects can watch the videos repeatedly to better evaluate the differences. For each video triplet, the subject is asked to choose which inpainting video is preferred. 44 subjects participated in our user study. Each participant watched averagely 20 triplets. Therefore, each pair of methods is compared about 220 times.

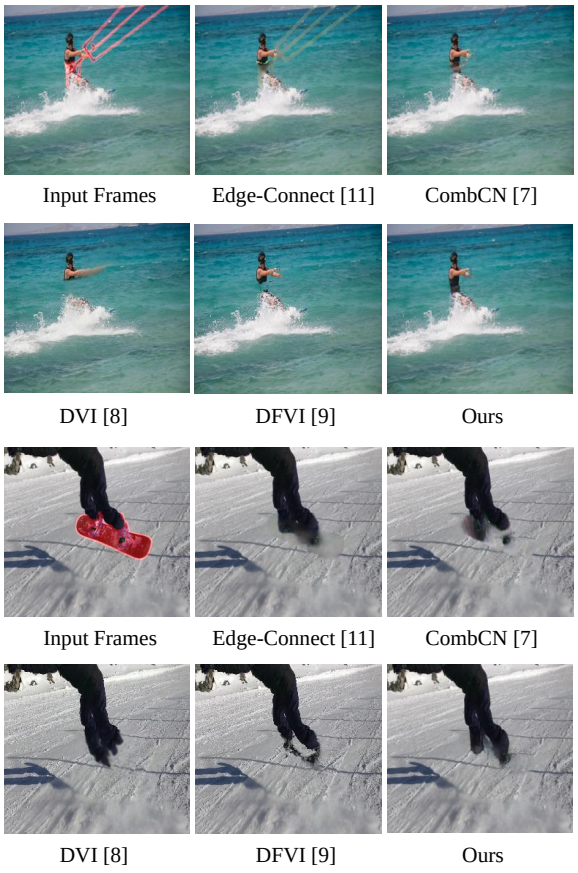

Fig. 7. Results of object foreground removal. The red masks in input frames indicate the objects to be removed. Our method produces results with plausible structures and details.

The preference results in the user study are shown in Fig. 6. Comparing to Edge-Connect [12], CombCN [8], DVI [9], our results are preferred by a significantly larger portion of subjects. When comparing with the flow-guided method DFVI [10], our method is preferred by 55.24% of the tests. Notably, our method is much faster than DFVI. Fig. 7 shows two examples of object removal using different methods. We can see that the inpainted results generated by our methods are visually better than existing methods. Compared to the blurry contents in the results of Edge-Connect, CombCN, and DVI, our method produces sharp object boundaries and fine visual details. Notably, though the completed contents using DFVI have sharp edges, the global structure of the human bodies is corrupted. In comparison, our method achieves more

TABLE II
ABLATION STUDIES ON YOUTUBEVOS. STRUCTURE INFERENCE, STRUCTURE ATTENTION MODEL, AND FLOW CONSTRAINED LOSS ARE DEMONSTRATED EFFECTIVE IN VIDEO INPAINTING.

	Fixed Square Mask			Moving Square Mask			Free-Form Mask			Inference
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID	Speed (fps)
TexNet	28.0174	0.9494	42.7164	33.8131	0.9705	8.2390	30.0680	0.9390	20.6358	7.6335
+Edge	29.5242	0.9520	36.2097	37.6630	0.9798	3.5161	33.8206	0.9659	6.6651	5.2356
+SAM	29.9918	0.9533	27.4198	38.2433	0.9807	2.5083	35.7783	0.9712	5.8786	5.1546
+Flow	**30.0590**	**0.9543**	**27.2431**	**38.8186**	**0.9824**	**2.3455**	**35.9613**	**0.9721**	**5.8694**	5.1546

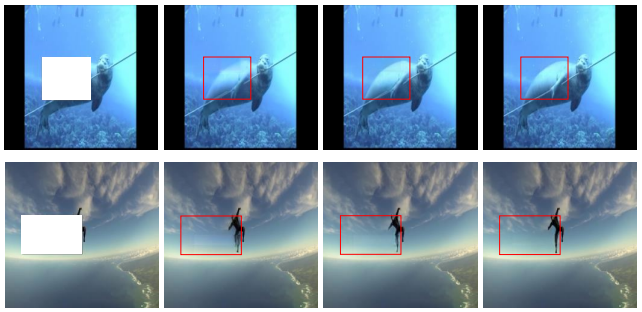

Fig. 8. Effects of structural guidance. By hallucinating edge maps first and then filling texture, we generate more completed and plausible target structure. Clearer boundaries can be obtained using the structure attention module.

intact and plausible structure with fine details. The results demonstrate the importance of utilizing structure information in video inpainting.

D. Ablation Study

To demonstrate the effectiveness of each component in our network, we conduct a series of ablation studies on the YouTubeVOS dataset with the first three mask settings. We test four variants of our model. The baseline model 'TexNet' only consists of the coarse-to-fine texture inpainting network without using edge maps as input and no SAM in the refinement module. This model simply integrates neighboring frames to predict the missing content for the current frame. Then we add the edge inpainting network ENet and feed the texture inpainting network with the completed edge maps to get the second model '+Edge'. The third model '+SAM' is constructed by adding the structure attention module on the second model. Finally, we add the flow constrained loss during training to get our full model '+Flow'. Especially, we only add the flow constraints in the training stage, which means it brings no computation costs to the inference process. The quantitative results are reported as in Table II.

1) Effect of Structure Clues: In Table II, The model '+Edge' brings large improvement over the baseline model. It indicates that sparse edges can provide effective structural guidance in video inpainting. When we further add SAM, extra improvement is obtained, demonstrating that the spatial correlation between edges and textures can be better embedded and absorbed by the texture inpainting network than simply feeding the hallucinated edge maps as extra channels into TexNet. The above analysis proves that the edge clues are effective guidance in video inpainting, which helps the network to predict more plausible frames with completed and detailed structure. Indeed, the structure inference module ENet brings extra time cost to the baseline TexNet from 7.6335 *fps* to 5.2356 *fps*. This is deserved because the inpainting quality is significantly improved.

TABLE III
COMPARISON OF PROPOSED STRUCTURE ATTENTION MODULE (SAM) AND SIMPLE 2-LAYER SPATIAL ATTENTION ON YOUTUBEVOS. SAM IS CAPABLE OF THE REVEALING POTENTIAL CORRELATION BETWEEN STRUCTURE INFORMATION AND VIDEO CONTENTS.

	Free-Form Mask		
	PSNR	SSIM	FID
+Edge	33.8206	0.9659	6.6651
+SimATT	34.4321	0.9685	6.3125
+SAM	**35.7783**	**0.9712**	**5.8786**

Fig. 8 shows the results generated using the three variants. It is obvious that after introducing structural guidance, the inpainted frames become more visually pleasing with sharper object boundaries. Besides, the edge maps predicted by our method are reasonable and clear, which well represent the image structure and show the strong edge inpainting ability of ENet. Thus, it is crucial to explore structural details in video inpainting.

2) Comparison of Proposed Structure Attention Module and Simple Attention: We propose a novel structure attention module (SAM) in TexNet to facilitate exploiting structure information of edge maps more effectively. This module is specifically designed for structural distilling in video inpainting. We conduct a comparison experiment between the proposed SAM and the commonly used simple 2-layer attention (SimATT) [41], *i.e.,* using a 2-layer convolution network to extract a spatial attention map from predicted edge maps, which is then applied to the same video feature as SAM. The model '+Edge' is used as the baseline. Results are shown in Table III. The performance of SAM is better than that of SimATT. It is because that the proposed SAM is more effective in revealing potential correlation between structure information in edge maps and video contents.

3) Visualization of Proposed Structure Attention Module: To further reveal the effects of structure attention module (SAM) in TexNet, some visualization results are given in Fig. 9. The red and green points in the input frames respectively indicate the selected missing pixels from object and background, of which the corresponding attention maps are given in Fig. 9 (c) and (d). Since most low-frequency missing textures have been well generated by the coarse inpainting network in TexNet, both the object and background

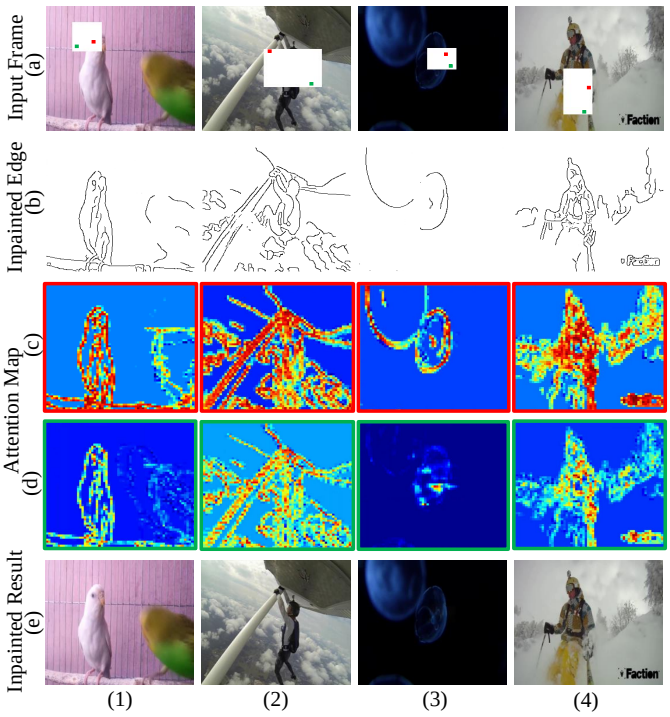

Fig. 9. Visualization of the attention map produced in proposed SAM. (a) is the input corrupted frame where the red and green points respectively indicate the selected missing pixels from object and background, of which the corresponding attention maps are given in Fig. 9 (c) and (d). (b) shows the corresponding completed edge map inpainted by ENet. Under the guidance of structural information, the final inpainted results are shown in (e).

pixels require high-frequency information around the edges to further enrich detailed textures. Especially, in term of the missing object pixels, attention maps in Fig. 9 (c) focus on the features of high-frequency edge textures as much as possible to preserve a consistent and clear object shape, which makes the results visually reasonable. In terms of attention map for the missing background pixels in Fig. 9 (d), SAM adaptively collects useful correlated information from different regions. For regions with smooth textures in the first and third example, SAM gives small weights for those edge regions. And for complex background inpainting in example (2) and (4), SAM also requires much high-frequency information. In summary, the above observations prove that it is reasonable to introduce the edge guidance into video inpainting, and the proposed structure attention module is an effective module in embedding the edge structure into the texture generation.

4) Effect of Flow for Temporal Coherence: We utilize temporal information to smoothen artificial flickers via two developed flow-guided warping losses during training. Table II shows that the quantitative performance is improved on all three mask settings by adding the flow guidance. Especially, we only add flow guidance in the training phase. Thus it brings performance gains without extra computation costs during testing. As Fig. 10 show, the synthesized contents in neighboring frames become more temporally consistent by adding the flow constrained losses. This proves that the proposed two flow-guided constraints in edge and texture inpainting networks are effective in enhancing temporal consistency.

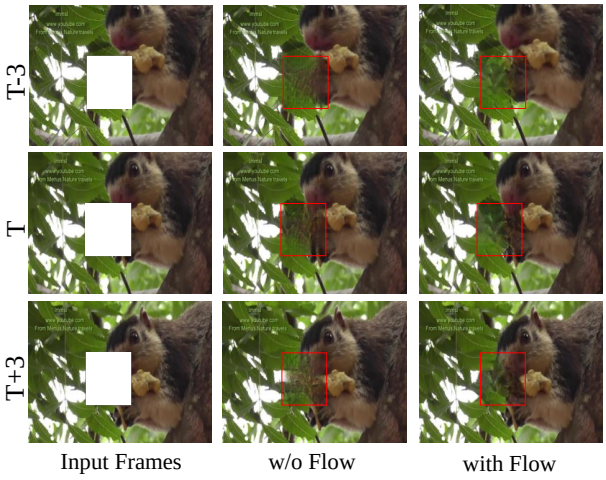

Fig. 10. Inpainting results of three neighboring frames. With the flow constraints during network training, the inpainting results are more temporally consistent without introducing image blurs.

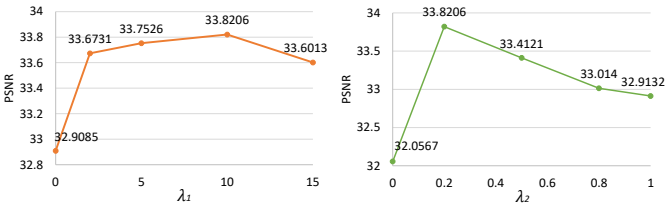

Fig. 11. Comparison of hyper parameters λ_1 and λ_2. The best result is obtained when $\lambda_1 = 10.0$ and $\lambda_2 = 0.2$.

5) Effects of Different Hyper Parameters: We conduct experiments to determine the hyper-parameters of λ_1 in Eq. (2) and λ_2 in Eq. (6). We use the integration model of ENet and TexNet in this experiment without SAM and flows. We first train ENet with different values of λ_1, and then train TexNet with restored edge maps while fixing ENet. The value of λ_2 is set as 0.2 when testing different λ_1. When determining λ_2 for TexNet, we set $\lambda_1 = 10.0$.

λ_1 is used in Eq. (2) as a weight of feature matching loss when training ENet. From the curve in Fig. 11, when increasing λ_1 from 0.0 to 2.0, the performance gain is obvious, which proves that the feature matching loss is effective in generating high-quality edge maps used for the final inpainted results. Then when λ_1 is increased from 2.0 to 10.0, slight improvement is obtained. The model obtains the best performance at $\lambda_1 = 10.0$. λ_2 in Eq. (6) is the weight of l_1-reconstruction loss of the coarse prediction in TexNet. When λ_2 is 0.0, the TexNet is trained without supervision of the coarse prediction, and the inpainting quality is significantly harmed. It demonstrates that the coarse-to-fine architecture is effective in TexNet. The best performance is obtained when $\lambda_2 = 0.2$. And performance drops when $\lambda_2 > 0.2$, which reflects that the supervision on the fine prediction networks is more important. Therefore, we set λ_2 as 0.2 in all the other experiments.

V. Conclusion

In this paper, we propose a novel structure-guided video inpainting approach which effectively utilizes structure information to recover fine-detailed content in corrupted videos. We first infer the target structure by predicting sparse edges in the missing region using an edge inpainting network. Then under the guidance of hallucinated edges, the missing content can be synthesized using the proposed coarse-to-fine texture network. The proposed structure attention module effectively exploits the correlation between structure and textures to improve the visual quality by producing more complete structure and visual details. Besides, the temporal coherence of the inpainting frames is further enhanced by our flow-assisted losses. Experiments on YouTubeVOS and DAVIS datasets under various mask settings demonstrate the effectiveness of our method on video inpainting and restoration tasks.

References

[1] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," *arXiv preprint arXiv:1806.03589*, 2018.

[2] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] K. A. Patwardhan, G. Sapiro, and M. Bertalmío, "Video inpainting under constrained camera motion," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 545–553, 2007.

[5] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 463–476, 2007.

[6] T. Shih, N. Tang, and J.-N. Hwang, "Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, pp. 347–360, 2009.

[7] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, 2014.

[8] C. Wang, H. Huang, X. Han, and J. Wang, "Video inpainting by jointly learning temporal structure and spatial details," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[9] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[10] R. Xu, X. Li, B. Zhou, and C. C. Loy, "Deep flow-guided video inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[11] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep blind video decaptioning by temporal aggregation and recurrence," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision Workshops*, 2019.

[13] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[14] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[15] G. Sridevi and S. S. Kumar, "Image inpainting based on fractional-order nonlinear diffusion for image reconstruction," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3802–3817, 2019.

[16] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 882–889, 2003.

[17] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.

[18] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics (ToG)*, vol. 28, no. 3, p. 24, 2009.

[19] K. Sangeetha, P. Sengottuvelan, and E. Balamurugan, "Combined structure and texture image inpainting algorithm for natural scene image completion," *Journal of Information Engineering and Applications*, vol. 1, no. 1, pp. 7–12, 2011.

[20] T. Ružić and A. Pižurica, "Context-aware patch-based image inpainting using markov random field modeling," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 444–456, 2014.

[21] D. Ding, S. Ram, and J. J. Rodríguez, "Image inpainting using nonlocal texture matching and nonlinear filtering," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1705–1719, 2019.

[22] Q. Guo, S. Gao, X. Zhang, Y. Yin, and C. Zhang, "Patch-based image inpainting via two-stage low rank approximation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 6, pp. 2023–2036, 2018.

[23] H. Lu, Q. Liu, M. Zhang, Y. Wang, and X. Deng, "Gradient-based low rank method and its application in image inpainting," *Multimedia Tools and Applications*, vol. 77, no. 5, pp. 5969–5993, 2018.

[24] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 29, no. 3, pp. 463–476, 2007.

[25] Y. Umeda and K. Arakawa, "Removal of film scratches using exemplar-based inpainting with directional median filter," in *2012 International Symposium on Communications and Information Technologies (ISCIT)*, 2012.

[26] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Temporally coherent completion of dynamic video," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 6, pp. 1–11, 2016.

[27] A. Ghanbari and M. Soryani, "Contour-based video inpainting," in *2011 7th Iranian Conference on Machine Vision and Image Processing*, 2011.

[28] A. Xia, Y. Gui, L. Yao, L. Ma, and X. Lin, "Exemplar-based object removal in video using gmm," in *2011 International Conference on Multimedia and Signal Processing*, 2011.

[29] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, 2012.

[30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[31] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[32] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.

[33] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[34] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision*, 2018.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[37] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "YouTube-VOS: Sequence-to-sequence video object segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018.

[38] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 DAVIS challenge on video object segmentation," *arXiv:1704.00675*, 2017.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017.
[41] S. Min, X. Chen, Z.-J. Zha, F. Wu, and Y. Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.