Structure-Guided Deep Video Inpainting

Chaoqun Wang, Xuejin Chen, *Member, IEEE*, Shaobo Min, Zheng-Jun Zha, *Member, IEEE*, and Jiaping Wang, *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 temporal 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 introduces temporal structure coherence to improve video inpanting results. In contrast to directly synthesizing the missing pixels, we first complete the edges in the missing regions to well represent scene structure 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 edge structure. Specifically, the SAM can model the semantic correlation between video textures and structural edges to generate structure-consistent inpainted videos. Besides, the flow information is used as a temporal consistency constraint 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 in 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 assist 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 by exploiting complementary information across neighboring frames and compositing visually pleasing content in the missing regions via patches [4], [5], [6]. 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 [7], [8], [9], [10]. A straightforward solution is to utilize 3D convolution layers to extract spatio-temporal features and predict missing contents with smooth motion [7]. To obtain temporally smooth results, contextual information from neighboring frames is aggregated to synthesize corrupted regions [8], [10], and optical flow is utilized in [9] to propagate pixel colors to unknown regions. By introducing motion guidance, these methods pay more attention to temporal smoothness; however, structure rationality and object details have not been well explored. Without definite representation and generation of the target image structures, these methods tend to produce over-smoothed regions. Similar observations have been obtained in image inpainting [2], [11]. To solve this problem, two-step methods have been proposed to complete object contours [2] or edge maps [11] first as auxiliary information to guide texture synthesis in image inpainting later. However, when applying these edge-first image inpainting methods to video inpainting, it brings another challenge in generating temporally coherent structures when human vision is significantly sensitive to temporal discontinuity that frequently occurs at edges.

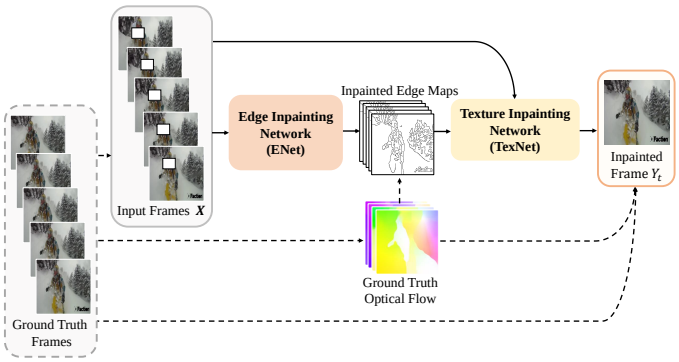

Fig. 1. Overview of our structure-guided inpainting network. We first complete the missing edges by aggregating information from neighboring frames to represent the target structure using the ENet. Then, the TexNet synthesizes the missing textures under the guidance of structure edges. Besides, the ground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

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 the motion information, we explore the correlation among structure, motion, and texture to complete the missing region with reasonable structure, rich visual details, and temporal coherence, as shown in Fig. 1. First, we design an edge

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.

inpainting network (ENet) to predict sparse edges in the missing region that represent the target structure information for each frame by exploiting the spatio-temporal neighboring information from adjacent frames. Then, under the guidance of completed edges, the texture inpainting network (TexNet) fills the textures via a coarse-to-fine architecture and a structure attention module (SAM). Specifically, the SAM is designed to guide the texture generation of TexNet by capturing the latent spatial relevance between video textures and completed structural edges. Notably, such a structure-texture relevance can effectively improve the fine-detailed frame generation in TexNet with fewer cracks and better object contours. 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 constraints, *i.e.,* the flow-guided edge warping loss and frame warping loss. Consequently, the inpainted frames using our approach are not only temporally consistent, but also completed in structure and rich in visual details.

Experiments on the YouTubeVOS and DAVIS datasets show that the proposed method obtains new state-of-the-art inpainting performance with low time cost. Our technical contributions can be summarized as follows:

- We propose a novel structure-guided video inpainting method, which integrates scene structure, texture, and motion to complete the missing region with valid structure, rich visual details, and temporal coherence.
- A structure attention module is designed to capture the correlation between structure 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 consistency of both 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 [12], [13] gradually propagate contents from surrounding areas to the missing region. Li *et al.* [14] attempt to solve the problem of localization of diffusion-based inpainted regions. Li *et al.* [15] define diffusion coefficients according to the relation between the damaged pixels and neighborhood pixels. Fractional-order nonlinear diffusion driven by difference curvature is proposed to produce clearer image details [16]. 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 [17], [18], are more widely studied. They formulate the completion task as a patch-based optimization problem. Barnes *et al.* [19] employ approximate nearest neighbor algorithm to fill the damaged regions. Sangeetha *et al.* [20] propose to propagate both linear structure and two-dimensional texture into the target region. Ružić *et al.* [21] introduce Markov random field to help search the most matched candidates. Ding *et al.* [22] 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.* [23] propose a simple two-stage low rank approximation to recover the corrupted region, which avoids time-consuming iterations. Lu *et al.* [24] adopt gradient-based low rank approximation. Patch-based image inpainting methods fill the missing content by borrowing and aggregating the most similar patches based on low-level image features from known regions and pasting them to unknown parts. However, this type of method usually fails when there is insufficient information in known regions or image textures are too complicated.

b) Patch-Based Video Inpainting: Patch-based video inpainting methods search similar patches and borrow appearances from known regions across frames. Newson *et al.* [6] extend the 2D PatchMatch algorithm [19] into 3D version to improve inpainting quality. Huang *et al.* [25] propose jointly estimating optical flow and textures to promote temporal coherence. Wexler *et al.* [26] constrain masked regions to synthesize coherent structures with respect to reference examples based on local structures. Umeda *et al.* [27] propose using directional median filter as complementation of patch-based filling. Some 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.* [28] first separate the two parts in videos, and fills the two parts accordingly with the help of contours. Xia *et al.* [29] make use of Gaussian mixture models to also distinguish moving foreground and still background, and process them separately.

However, the patch-searching process makes patch-based video inpainting methods suffer from high computational complexity, which limits 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. The convolution neural network (CNN) is first introduced for image denoising and inpainting in [30], where CNN is used to directly synthesize image contents in the masked regions. To improve the photorealism of the completed results, a generative adversarial network is employed [31]. Then, Yang *et al.* [32] take advantages of multi-scale representation to boost details generation. Multiple discriminators are used to constrain both global and local coherence of image contents [33]. Yu *et al.* [34] propose the 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 [35] and employing gated convolution [1] for dynamic feature selection. Both these two methods want to handle image inpainting with irregular regions, which is hard for vanilla convolutions. While these methods tend to generate

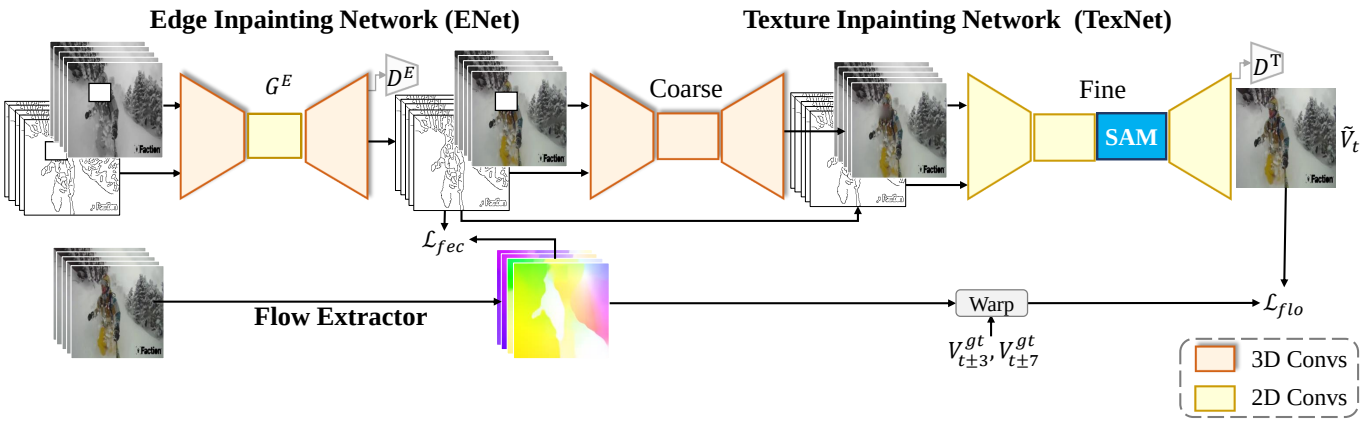

Fig. 2. The detailed architecture of our network. ENet adopts an encoder-decoder architecture to complete edges in the missing regions. TexNet utilizes a coarse-to-fine manner to inpaint the final frames. Ground truth optical flow between adjacent frames is extracted for edge consistency loss and frame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

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 [11]. 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/contour-based methods, simply extending it from image inpainting to video inpainting by 3D convolutions inevitably fails because of no guarantee on the temporal coherence. 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.

d) Deep Video Inpainting: Besides the challenge in maintaining temporal coherence, image inpainting methods do not fully utilize useful complementary information in neighboring frames, which could help large hole completion in videos. Several methods regarding video inpainting based on deep neural networks have been proposed recently. The first deep-learning-based video inpainting method is CombCN [7], 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 contents in the current frame [8], [10]. Instead of filling pixel colors directly using CNNs, a deep flow completion network is first proposed to estimate optical flow in the missing region and then pixel colors are propagated based on it [9]. 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 when 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 introduce a structure attention mechanism. Under the structural guidance, more visually pleasing results could be synthesized with reasonable 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 complete a target frame $\widetilde{V}_t$ at time t, we take total T 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.

As Fig. 2 shows, our framework consists of three main components. The first part is an edge inpainting network (ENet) that targets to recover the missing edges, and the second part is a coarse-to-fine texture inpainting network (TexNet) that aims to complete the missing appearance details under the structure guidance. The third part leverages the ground truth motion flow as auxiliary constraints to preserve temporal coherence of both completed edge maps and textures during the training stage.

A. Edge Inpainting Network

The edge inpainting network (ENet) complete the edges in the missing region to depict object shapes, which is much easier to hallucinate without complex colors. Therefore, under the guidance of well-completed object structure, we can better inpaint the whole texture.

Given the input corrupted frames $\boldsymbol{V}$ as well as the corresponding binary masks $\boldsymbol{M}$, where $M_t^p = 1$ indicating corrupted pixel on position p in the missing regions at time t. A Canny edge detector is first used to extract the corresponding edge maps $\boldsymbol{E}^i$. The input of ENet consists of the incomplete grayscale version of frames $\boldsymbol{V}^g$, initial edge maps $\boldsymbol{E}^i = \boldsymbol{E}^{gt} \odot \big(1 - \boldsymbol{M}\big)$, and their corresponding masks $\boldsymbol{M}$. As shown in Fig. 2, ENet is denoted by G^E, which 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 the 3-dimensional convolution operations. The intermediate 2D residual blocks are used to enlarge the spatial receptive fields by using 2-dimensional convolutions of large kernel size. Such a 3D+2D

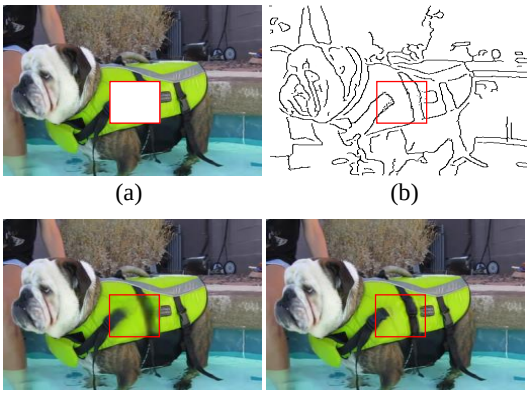

Fig. 3. To inpaint missing regions in the target corrupted frame (a), our ENet first completes corresponding sparse edge (b), which well represents the structure of the missing contents. Then TexNet progressively replenishes textures under the guidance of synthesized edges from coarse (c) to fine (d).

convolution architecture can achieve a good trade-off between spatial and temporal perception. The inpainted T edge maps 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:

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

where the discriminator D^E follows the 70×70 PatchGAN architecture [36]. $\mathcal{L}^E_{adv}$ and $\mathcal{L}^E_{fm}$ are the adversarial loss and feature matching loss. 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 predicted edge maps by:

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

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

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

where D^E_k is the output of the k-th layer in L-layer D^E, while N_k is the element number 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.

Consequently, the 3D+2D architecture and two-level loss function enable the ENet to inpaint the missing structural edges accurately. An example of the generated edge map is given in Fig. 3 (a) and (b).

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, *i.e.,* TexNet. Notably, the object structure, *e.g.,* object shapes, has been captured in the completed edge frames $\widetilde{\boldsymbol{E}}$. Thus, it becomes much easier to fill the missing texture with the structural guidance of $\widetilde{\boldsymbol{E}}$.

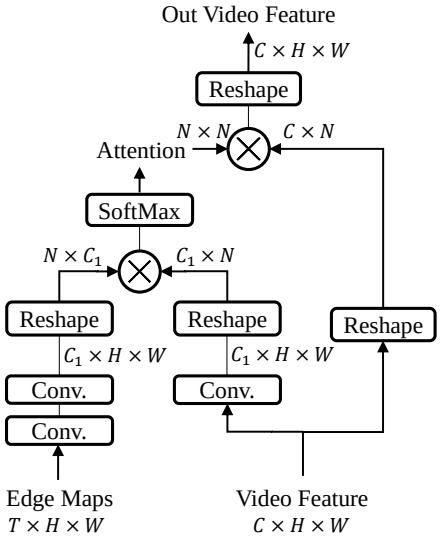

Fig. 4. Architecture of the structure attention module. C is the channel of the input video features, and $N = H \times W$. $\otimes$ represents matrix multiplication. Usually, we set $C_1 = C/8$.

To synthesize realistic frame texture, the proposed TexNet adopts a coarse-to-fine architecture, as shown in Fig. 2. Specifically, 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, which targets to produce 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 the coarse 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, which is used as an intermediate layer in the refinement network in Fig. 2. 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, which represents the correlation between the structure and high-level 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, the inpainted content by TexNet becomes more realistic, as Fig. 3 (d) shows.

Specifically, the coarse inpainting network and the refine-

ment network are trained end-to-end by:

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

where G^T denotes the TexNet and D^T is the discriminator. Inspired by [11], 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:

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

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

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

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

From the results in Fig. 3 (c) and (d), it can be seen that the refinement network can exactly refine the inpainting results with more clear contours and textures, and the combined loss of $\mathcal{L}^T_{rec}$ and $\mathcal{L}^T_{adv}$ can well train the coarse and refinement networks jointly.

C. Flow-Guided Temporal Coherence Enhancement

Besides the structural guidance, the motion information is also considered to maintain temporal consistency among frames. To this end, the optical flow is employed in the training stage of both ENet and TexNet.

During training, a set of flow maps $\boldsymbol{O}$ between the current frame V^{gt}_t and its neighboring frames are generated using a pre-trained flow extraction network, such as FlowNet2.0 [37]. Specifically, $\boldsymbol{O}$ consists of four flow maps $(O_{t\Rightarrow t-7}, O_{t\Rightarrow t-3}, O_{t\Rightarrow t+3}, O_{t\Rightarrow t+7})$. $\boldsymbol{O}$ contains the ground truth motion information between frames, which is important to both edge and texture generation.

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:

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

where $\phi(O_{t\Rightarrow t+k},E^{gt}_{t+k})$ is the warping operation which warps the edge map E^{gt}_{t+k} 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 $\mathcal{L}^E_{fec}$, the loss function of Eq. (2) for ENet becomes:

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

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

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

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

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

After adding the motion information to both the training process of ENet and TexNet, the temporal consistency can be preserved 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 [38] and DAVIS [39], 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 region are fixed through the whole video.
- (2) Moving square mask: The position and size of the square mask change over frames.
- (3) Free-from mask: We apply irregular masks which imitate hand-drawn masks on each frame, following [35].
- (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 validating, and 508 for testing. Since YoutubeVOS has no dense foreground mask annotations, we only use it for evaluation of mask settings (1), (2), and (3). DAVIS dataset contains 90 video sequences that are annotated with foreground object masks 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 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. Learning rate is set as $1e-4$ for G^T, and $4e-4$ for D^T. We first train ENet and

TABLE I
COMPARISONS WITH FOUR STATE-OF-THE-ART METHODS PROPOSED IN 2019 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 [11]	28.6446	0.9484	38.2116	30.7478	0.9647	16.2739	25.6693	0.9088	43.0366	22.81
CombCN [7]	27.9668	0.9515	40.7199	31.5776	0.9678	13.8383	32.1862	0.9626	19.1191	8.1634
DVI [8]	28.0846	0.9468	39.9377	36.8598	0.9728	7.2315	33.5549	0.9646	9.3797	1.2275
DFVI [9]	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. 5. Visualization comparison on YouTubeVOS with state-of-the-art methods. Our method produces results with more complete object structures and finer details.

TexNet without corresponding flow losses, and then add the flow losses to refine the final results when flow assistance is utilized. Adam optimizer with $\beta = (0.9, 0.999)$ is used for all sub-networks. We do not use weight decay in training. As for the hyper-parameters, $\lambda_1 = 10.0, \lambda_2 = 0.2$.

Different data preparations and evaluation metrics are used according to mask settings. a) We randomly generate masks for training videos in terms of mask settings (1), (2), and (3). Masked videos are used for testing. Three commonly-used metrics, including structural similarity index (SSIM) [40], peak signal-to-noise ratio (PSNR), and Fréchet Inception Distance (FID) [41] are used to quantitatively evaluate the performance of our method. b) For the mask setting (4), we prepare masks of random shapes and motions to synthesize the foreground object masks in the training stage. The network is first trained on the YoutubeVOS dataset and then finetuned on DAVIS. Since there is no ground truth available for this setting, which means that we can not use quantitive evaluations to measure the output quality, we conduct a user study for video foreground object removal.

B. Main Results on Video Inpainting

We compare the proposed method with four state-of-the-art video inpainting methods [11], [7], [8], [9] for the first three mask settings on the YouTubeVOS dataset. We train [11], [9] using their published codes and re-implement [7] according to their paper. As for [8], we use the officially provided model since there are no available training codes.

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 [8] and nine times faster than DFVI [9]. We also notice that all models have different performances with different masks. The best performance of an individual model is obtained when using moving square masks, which is because the network can learn complementary information from neighboring frames. 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 2D image inpainting method, Edge-Connect [11], which also predicts edges to represent the structure, our method greatly increases the completion performance by leveraging neighboring frames to complete edges and synthesize textures. Thus, our method can generate more temporal coherent and realistic contents. The inference speed of Edge-Connect is fast, the reason of which is that it does not consider axillary temporal information between frames.

Compared with the second-best video inpainting method DFVI [9], our method can produce 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 and also indicates that structural clues can bring strong promotion to video inpainting.

C. Results on Object Removal

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 [11], [7], [8], [9]. 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.

The preference results in the user study are shown in Fig. 6. Comparing to Edge-Connect [11], CombCN [7], DVI [8], our results are preferred by a significantly larger portion of subjects. When comparing with the flow-guided method DFVI [9], 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 intact and plausible structure with fine details. The results demonstrate the importance of utilizing structure information in video inpainting.

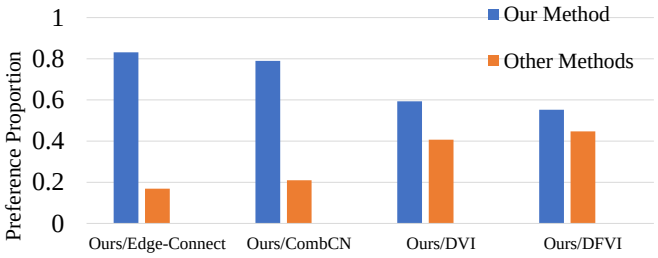

Fig. 6. Results of user study. Ours are preferred by more participants compared to state-of-the-art methods.

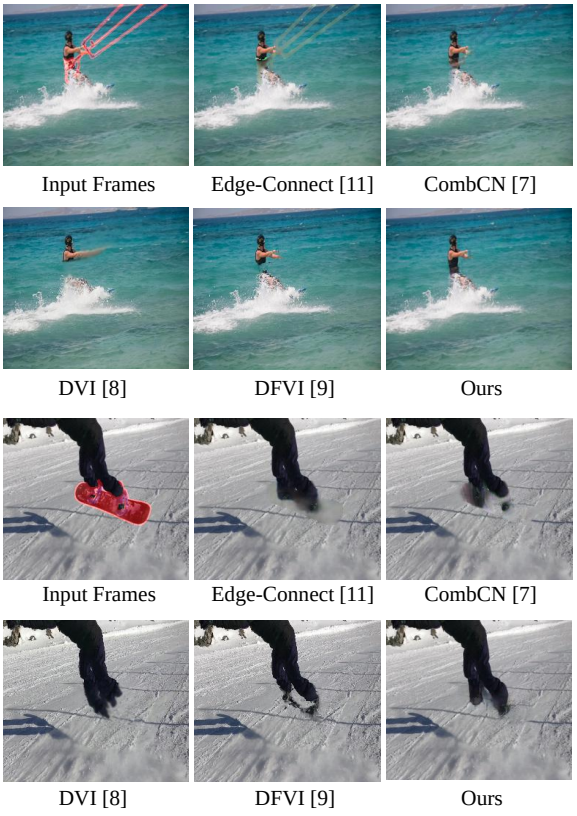

Fig. 7. Results of object foreground removal. The red masks in input frames indicate the objects to be removed. Our method produces results with plausible structures and details.

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

TABLE II
ABLATION STUDIES ON YOUTUBEVOS. STRUCTURE INPUT, STRUCTURE ATTENTION MECHANISM, AND FLOW ASSISTANCE 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
Ours	**30.0590**	**0.9543**	**27.2431**	**38.8186**	**0.9824**	**2.3455**	**35.9613**	**0.9721**	**5.8694**	5.1546

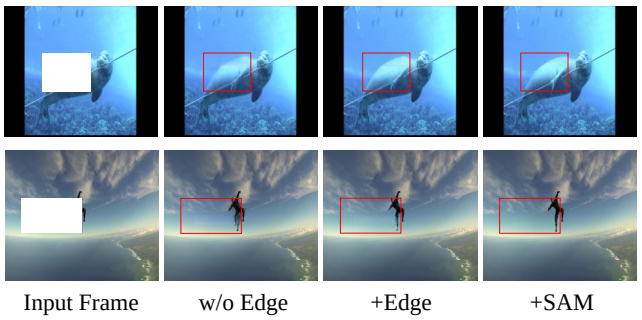

Fig. 8. Effects of structure guidance. By inpainting edge maps first and then filling texture, we generate more completed target structure. Clearer boundaries can be obtained using the structure attention module.

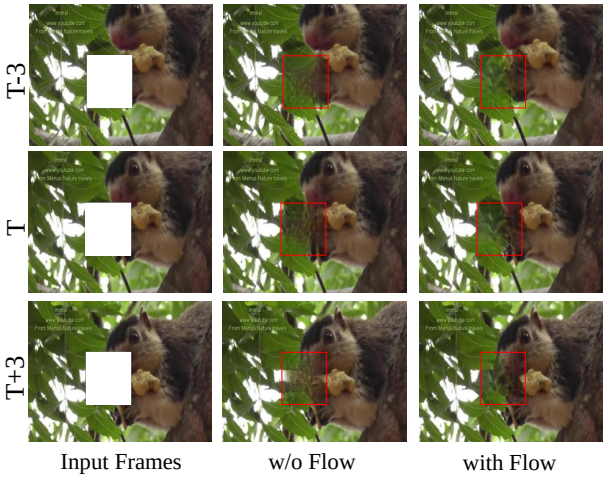

Fig. 9. Inpainting results of three neighboring frames. With the flow assistance, the inpainting results are more temporally consistent without introducing image blurs.

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 and fed 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 guidance to get our full model 'Ours'. Especially, we only add the flow guidance in the training stage, which means it brings no computation costs to test. The quantitative results are reported as in Table II.

1) Effect of Structure Clues: In Table II, '+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 to '+Edge', 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 completed 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 accurate frames. The inpainted edge map in Fig. 3 shows that ENet is capable of generating completed and detailed structure. Indeed, the structure 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.

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. In this part, we conduct a comparison experiment between the proposed SAM and the commonly used simple 2-layer attention (SimATT) [42], *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 capable of revealing potential correlation between structure information in edge maps and video contents. Thus it is easier for TexNet to utilize structure information to obtain better results.

3) Effect of Flow for Temporal Coherence: In our method, we utilize temporal information to smoothen artificial flickers via two developed flow-guided warping losses. From Table II, we can see that the quantitative performance is improved on all three settings by adding the flow guidance. Especially,

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**

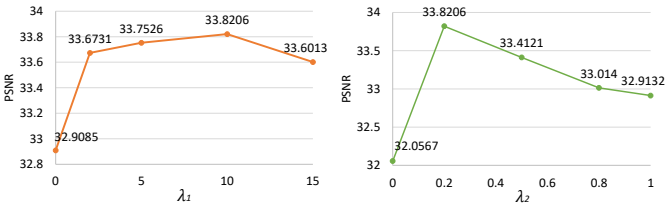

Fig. 10. Results of hyper parameters λ_1 and λ_2. The best result is obtained when $\lambda_1 = 10.0$ and $\lambda_2 = 0.2$.

we only add flow guidance in the training phase, so it can bring gains without extra computation costs during testing. The results show that the improvement from the optical flow is smaller than that from the structural guidance. The reason is that the flow is only used in the training stage as temporal guidance, while the edge is used during the inference. It should be noted that predicting the completed optical flow among several frames takes more time than the edges during inference. Thus, we use the flow in training and the structure edge in both training and testing, which can achieve a good balance between the quality improvement and the inference efficiency.

As shown in Fig. 9, the synthesized contents in neighboring frames become more temporally consistent after adding the flow guidance. This proves that the proposed two flow-guided constraints in edge and texture inpainting networks are effective in preserving the temporal consistency.

4) Effects of Different Value of Hyper Parameters: In this part, we conduct experiments to determine the hyper-parameters of λ_1 in Eq. (2) and λ_2 in Eq. (6). We use edge inputs for TexNet in this part, without SAM and flow assistance. When testing λ_1, we first train ENet with different values of λ_1, and then train TexNet with generated edge maps by fixed ENet. The value of λ_2 is set as 0.2. When determining λ_2, we keep the value of λ_1 as 10.0, then train TexNet with different λ_2.

λ_1 is used in Eq. (2), which is used to denote the weight of feature matching loss when training ENet. From the results in Fig. 10, 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. Finally, when $\lambda_1 = 10.0$, the model obtains the best performance. Therefore, we set λ_1 to 10.0 for the experimental settings.

As for λ_2 in Eq. (6), which 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 l_1-reconstruction loss of the coarse prediction, the result is heavily harmed. It demonstrates that the coarse-to-fine architecture is effective in TexNet. The best performance of the result is obtained when λ_2 is set as 0.2. And performance drops when $\lambda_2 > 0.2$, which reflects that the constraints on fine predictions are more important than that on coarse ones. So we set λ_2 as 0.2 in experiments.

V. CONCLUSION

In this paper, we propose a novel structure-guided video inpainting approach, which effectively utilizes structure information to generate fine-detailed frames. The edges in the missing region are first estimated to represent the target structure explicitly, using an edge inpainting network. Then under the guidance of the sparse edges, the missing content can be synthesized with intact structure and fine details. Our proposed structure attention module effectively exploits the correlation between structure and textures to improve visual quality. Besides, the temporal coherence of the inpainting frames is further enhanced by our flow-assisted losses. Experiments on YouTubeVOS and DAVIS datasets demonstrate the effectiveness of our method.

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, year=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] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: Structure guided image inpainting using edge prediction," pp. 0–0, 2019.
[12] 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, pp. 417–424.
[13] 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.
[14] H. Li, W. Luo, and J. Huang, "Localization of diffusion-based inpainting in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 3050–3064, 2017.

[15] K. Li, Y. Wei, Z. Yang, and W. Wei, "Image inpainting algorithm based on tv model and evolutionary algorithm," *Soft Computing*, vol. 20, no. 3, pp. 885–893, 2016.
[16] G. Sridevi and S. S. Kumar, "Image inpainting based on fractional-order nonlinear diffusion for image reconstruction," *Circuits, Systems, and Signal Processing*, pp. 1–16, 2019.
[17] 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.
[18] 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.
[19] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," in *ACM Transactions on Graphics (ToG)*, vol. 28, no. 3, 2009, p. 24.
[20] 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.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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, p. 196, 2016.
[26] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 463–476, 2007.
[27] 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)*. IEEE, 2012, pp. 6–11.
[28] A. Ghanbari and M. Soryani, "Contour-based video inpainting," in *2011 7th Iranian Conference on Machine Vision and Image Processing*. IEEE, 2011, pp. 1–5.
[29] 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*, vol. 1. IEEE, 2011, pp. 366–370.
[30] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, 2012.
[31] 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.
[32] 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, pp. 6721–6729.
[33] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 107, 2017.
[34] 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, pp. 5505–5514.
[35] 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 *The European Conference on Computer Vision*, 2018.
[36] 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.
[37] 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.
[38] 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 (ECCV)*, 2018, pp. 585–601.
[39] 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.
[40] 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.
[41] 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, pp. 6626–6637.
[42] 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*, vol. 33, 2019, pp. 4578–4585.