Deep Stereo Video Inpainting

Zhiliang Wu[1], Changchang Sun[2], Hanyu Xuan[3], Yan Yan[2*]
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] Department of Computer Science, Illinois Institute of Technology, USA
[3] School of Big Data and Statistics, Anhui University, China

Abstract

Stereo video inpainting aims to fill the missing regions on the left and right views of the stereo video with plausible content simultaneously. Compared with the single video inpainting that has achieved promising results using deep convolutional neural networks, inpainting the missing regions of stereo video has not been thoroughly explored. In essence, apart from the spatial and temporal consistency that single video inpainting needs to achieve, another key challenge for stereo video inpainting is to maintain the stereo consistency between left and right views and hence alleviate the 3D fatigue for viewers. In this paper, we propose a novel deep stereo video inpainting network named SVINet, which is the first attempt for stereo video inpainting task utilizing deep convolutional neural networks. SVINet first utilizes a self-supervised flow-guided deformable temporal alignment module to align the features on the left and right view branches, respectively. Then, the aligned features are fed into a shared adaptive feature aggregation module to generate missing contents of their respective branches. Finally, the parallax attention module (PAM) that uses the cross-view information to consider the significant stereo correlation is introduced to fuse the completed features of left and right views. Furthermore, we develop a stereo consistency loss to regularize the trained parameters, so that our model is able to yield high-quality stereo video inpainting results with better stereo consistency. Experimental results demonstrate that our SVINet outperforms state-of-the-art single video inpainting models.

1. Introduction

Video inpainting aims to fill in missing region with plausible and coherent contents for all video frames. As a fundamental task in computer vision, video inpainting is usually adopted to enhance visual quality. It has great value in many practical applications, such as scratch restoration [2], undesired object removal [34], and autonomous driving [24]. In recent years, relying on the powerful features extraction capabilities of convolutional neural network (CNN), existing deep single video inpainting methods [6, 13, 15, 18, 20, 23, 42, 46] have shown great success. With the development of augmented reality (AR), virtual reality (VR) devices, dual-lens smartphones, and autonomous robots, there is an increasing demand for various stereo video processing techniques, including stereo video inpainting. For example, in some scenarios, we not only remove objects and edit contents, but also expect to recover the missing regions in the stereo video. Although the traditional stereo video inpainting methods [31, 32] based on patch optimization have been preliminarily studied, the stereo video inpainting based on deep learning has not been explored.

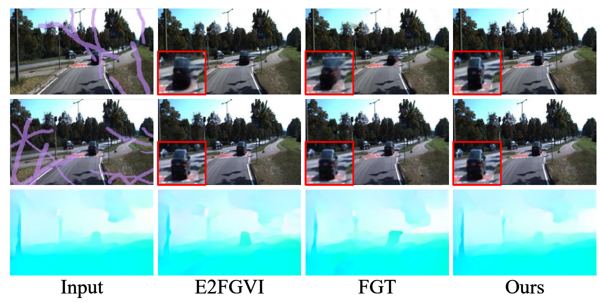

Figure 1. An example of visual comparison with state-of-the-art single video inpainting models (E2FGVI [23] and FGT [48]) on stereo video inpainting. As shown here, directly using the single video inpainting method to generate missing contents on the left view (first row) and right view (second row) will lead to severe stereo inconsistency. In contrast, the proposed method not only generates vivid textures, but also the parallax flow (third row) between two views is closer to the ground-truth (third row of the input column). The closer the parallax flow is to ground-truth, the better the stereo consistency is maintained.

A naive solution of stereo video inpainting is to directly apply the single video inpainting methods by completing the missing regions of left and right views, respectively. However, inpainting an individual video that only considers the undamaged spatial-temporal statistics of one view will

*Corresponding author

ignore the geometric relationship between two views, causing severe stereo inconsistency as shown in Fig. 1. Besides, another way to solve this task is process the stereo video frame-by-frame using the stereo image inpainting methods. For example, Li et al. [22] designed a Geometry-Aware Attention (GAA) module to learn the geometry-aware guidance from one view to another, so as to make the corresponding regions in the inpainted stereo images consistent. Nevertheless, compared to its image counterpart, stereo video inpainting still needs to concern the temporal consistency. In this way, satisfactory performance cannot be achieved by extending stereo image inpainting technique to stereo video inpainting task. Therefore, to maintain temporal and stereo consistency simultaneously, there are two key points need to be considered: (i) temporal modeling between consecutive frames (ii) correlation modeling between left view and right view.

In fact, on the one hand, the missing contents in one frame may exist in neighboring (reference) frames of a video sequence. Thus, the temporal information between the consecutive frames can be explored to generate missing contents of the current (target) frame. For example, a classical technology pipeline is "alignment–aggregation", that is, the reference frame is first aligned to eliminate image changes between the reference frame and target frame, and then the aligned reference frame is aggregated to generate the missing contents of the target frame. On the other hand, correlation modeling between two views has been studied extensively in the stereo image super-resolution task [3, 39, 44]. For instance, Wang et al. [39] proposed the parallax attention module (PAM) to tackle the varying parallax problem in the parallax attention stereo super-resolution network (PASSRnet). Ying et al. [44] developed a stereo attention module (SAM) to address the information incorporation issue in the stereo image super-resolution models. More recently, Chen et al. [3] designed a cross-parallax attention module (CPAM) which can capture the stereo correspondence of respective additional information.

Motivated by above observation and analysis, in this paper, we propose a stereo video inpainting network, named SVINet. Specifically, SVINet first utilizes a self-supervised flow-guided deformable temporal alignment module to align the reference frames on the left and right view branches at the feature level, respectively. Such operation can eliminate the negative effect of image changes caused by camera or object motion. Then, the aligned reference frame features are fed into a shared adaptive feature aggregation module to generate missing contents of their respective branches. Note that the missing contents of one view may also exist in another view, we also introduce the most relevant target frame from another view when completing the missing regions of the current view, which can avoid the computational complexity problem caused by simply aggregating all video frames. Finally, a modified PAM is used to model the stereo correlation between the completed features of the left and right views. Beyond that, inspired by the success of end-point error (EPE) [10] in optical flow estimation [11], we introduce a new stereo consistency loss to regularize training parameters, so that our model is able to yield high-quality stereo video inpainting results with better stereo consistency. We conduct extensive experiments on two benchmark datasets, and the experimental results show that our SVINet surpasses the performance of recent single video inpainting methods in the stereo video inpainting.

To sum up, our contributions are summarized as follows:

- We propose a novel end-to-end stereo video inpainting network named SVINet, where the spatially, temporally, and stereo consistent missing contents for corrupted stereo video are generated. To the best of our knowledge, this is the first work using deep learning to solve stereo video inpainting task.

- Inspired by the end-point error (EPE) [10], we design a stereo consistency loss to regularize training parameters of SVINet, so that the training model can improve the stereo consistency of the completed results.

- Experiments on two benchmark datasets demonstrate the superiority of our proposed method in both quantitative and qualitative evaluations. Notably, our method also shed light on the subsequent research of stereo video inpainting.

2. Related Works

Single Video Inpainting. With the rapid development of deep learning, several deep learning-based methods have been proposed for video inpainting and achieved significant results in terms of the inpainting quality and speed. These deep learning-based methods can be roughly classified into three lines: 3D convolution methods, optical flow methods, and attention ones. 3D convolution methods [2, 16, 26] usually reconstruct the missing contents by directly aggregating temporal information from neighbor frames through 3D temporal convolution. For example, Wang et al. [37] proposed the first deep learning-based video inpainting network, which consists of a 3D CNN for temporal prediction and a 2D CNN for spatial detail recovering. Further, Kim et al. [16] adopted a recurrent 3D-2D feed-forward network to aggregate the temporal information of the neighbor frames into missing regions of the target frame. However, 3D CNN has relatively higher computational complexities compared with 2D CNN, limiting the application of these methods. To alleviate this problem, some researchers treated the video inpainting as a pixel propagation problem and designed the video inpainting approaches [6, 14, 15, 23, 43, 49, 51] based

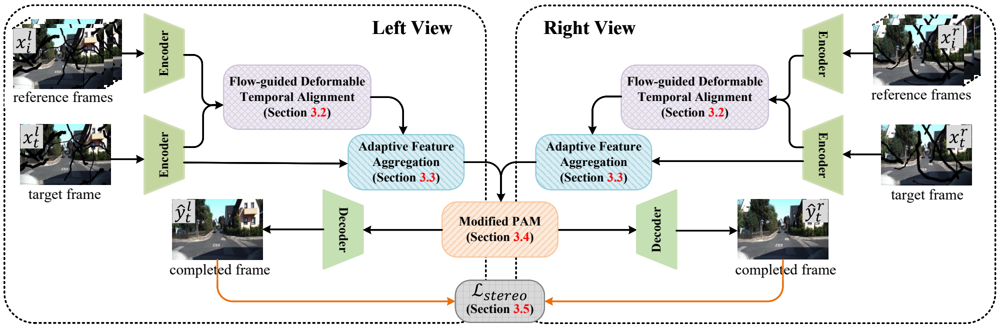

Figure 2. **Illustration of the proposed stereo video inpainting network (SVINet).** Flow-guided deformable temporal alignment module is used to align reference frame on the left and right view branches at the feature level, which aims to eliminate the effect of image changes caused by camera or object motion. Then, the aligned reference frame features are fed into adaptive feature aggregation module to generate missing contents of their respective branches. Finally, the completed features on the two branches are used to model the stereo consistency between the left and right views through the modified PAM. Furthermore, we also design a stereo consistency loss $\mathcal{L}_{stereo}$ to regularize the trained parameters, so that our model is able to yield video inpainting results with high-quality stereo consistency.

on optical flow. These methods first introduce a deep flow completion network to restore the flow sequence and then use the restored flow sequence to fill the relevant pixels of the missing regions of the neighbor frames. For instance, Xu et al. [43] used the flow field completed by a coarse-to-fine deep flow completion network to guide relevant pixels into the missing regions. Based on this, Gao et al. [6] further improved the performance of video inpainting by explicitly completing the flow edges. Zou et al. [51] corrected the spatial misalignment in the temporal feature propagation stage by the completed optical flow. Although have shown promising results, these methods fail to capture the visible contents of long-distance frames, and thus decrease the inpainting performance in the scene of large objects and slowly moving objects.

To effectively model the long-distance correspondence, the state-of-the-art methods [20,21,25,27,33–35,41,45,48] use the attention mechanism to capture long-term correspondences. In this way, the available content at distant frames can be globally propagated into missing regions. For example, Zeng et al. [45] proposed the first transformer model for video inpainting by learning a multi-layer multi-head transformers. Further, Liu et al. [25] improved edge details of missing contents by using soft split and soft composition operations in transformer. Ren et al. [33] developed a novel Discrete Latent Transformer (DLFormer) by formulating video inpainting task into the discrete latent space. In spite of these methods have achieved unprecedented performance in the single video inpainting task, the naive extension of these methods to stereo video inpainting tasks will lead to severe stereo inconsistency between two views.

Stereo Image/Video Inpainting. Stereo image inpainting is a sub-task of image inpainting, and several traditional methods have been proposed. Wang et al. [38] proposed a new stereo image inpainting algorithm for simultaneous color and depth inpainting. Hervieu et al. [9] used the complete disparity maps to fill in missing regions in a way that avoids the creation of 3D artifacts. However, due to the common limitation of conventional single image inpainting methods, they fail to generate meaningful structures when facing complex semantic scenes in the missing regions. Fortunately, the development of convolutional neural network brings new opportunities for stereo image inpainting. Chen et al. [4] designed the first end-to-end stereo image inpainting network based on the encoder-decoder structure. However, this method can only deal with square holes in the centre. Ma et al. [28] proposed SICNet for stereo image inpainting, which associates the two views by a feature map concatenation operation to ensure the stereo consistency of the completed results. Further, Li et al. [22] designed an Iterative Geometry-Aware Cross Guidance Network (IGC-Net), which performs inpainting on the stereo images by exploring and integrating the stereo geometry in an iterative manner. While these stereo image inpainting methods have achieved promising results, naively using these algorithms on individual stereo video frames to fill missing regions will lose inter-frame motion continuity, resulting in flicker artifacts in the inpainted video.

Compared to stereo image inpainting, stereo video inpainting presents an additional challenge in preserving temporal consistency. Traditional stereo video inpainting methods [31, 32] formulate the inpainting process as a patch-based optimization problem, *i.e.*, searching the similar patches from the known regions to synthesize missing contents, and using a view consistency constraint to ensure the stereo consistency of the results. Similar to traditional stereo image inpainting, these methods fail to complete scenes with complex semantics. Inspired by the success of deep learning in single video inpainting task, we propose the first deep stereo inpainting model in this paper, which provides a strong benchmark for subsequent research.

3. Method

3.1. Network Overview

Given a corrupted stereo video sequence $(\boldsymbol{X}^l, \boldsymbol{X}^r) = \{(\boldsymbol{x}_1^l, \boldsymbol{x}_1^r), (\boldsymbol{x}_2^l, \boldsymbol{x}_2^r), \ldots, (\boldsymbol{x}_T^l, \boldsymbol{x}_T^r)\}$ consisting of T frame pair, where $\boldsymbol{x}_i^l$ and $\boldsymbol{x}_i^r$ denote the i-th corrupted frames of the left and right stereo video $\boldsymbol{X}^l$ and $\boldsymbol{X}^r$, respectively. Let $(\boldsymbol{M}^l, \boldsymbol{M}^r) = \{(\boldsymbol{m}_1^l, \boldsymbol{m}_1^r), (\boldsymbol{m}_2^l, \boldsymbol{m}_2^r), \ldots, (\boldsymbol{m}_T^l, \boldsymbol{m}_T^r)\}$ denote the corresponding frame-wise masks, which is used to indicate missing or corrupted regions. The goal of stereo video inpainting is to generate an inpainted stereo video sequence pair $(\widehat{\boldsymbol{Y}}^l, \widehat{\boldsymbol{Y}}^r) = \{(\widehat{\boldsymbol{y}}_1^l, \widehat{\boldsymbol{y}}_1^r), (\widehat{\boldsymbol{y}}_2^l, \widehat{\boldsymbol{y}}_2^r), \ldots, (\widehat{\boldsymbol{y}}_T^l, \widehat{\boldsymbol{y}}_T^r)\}$, which should be spatially, temporally, and stereo consistent with the original video sequence pair $(\boldsymbol{Y}^l, \boldsymbol{Y}^r) = \{(\boldsymbol{y}_1^l, \boldsymbol{y}_1^r), (\boldsymbol{y}_2^l, \boldsymbol{y}_2^r), \ldots, (\boldsymbol{y}_T^l, \boldsymbol{y}_T^r)\}$.

To achieve this goal, we propose a stereo video inpainting network named SVINet. As shown in Fig.2, SVINet consists of a frame-level encoder, a Flow-guided Deformable Temporal Alignment Module (FDTAM), an Adaptive Feature Aggregation Module (AFAM), a Parallax Attention Module (PAM) and a frame-level decoder. The frame-level encoder is built by stacking several 2D convolution layers, which aims at encoding deep features from low-level pixels of each frame. Similarly, the frame-level decoder is designed to decode inpainted features into frames. Besides, FDTAM, AFAM, and PAM are the core components of our proposed model. FDTAM performs reference frame alignment on the left and right view branches at the feature level, which aims to eliminate the effect of image changes caused by camera or object motion. After obtaining the aligned reference frame features, AFAM is used to generate missing contents of their respective branches. Note that the missing contents of one view may also exist in another view, so we also introduce video frames of another view when generating the missing contents of the current view. Finally, the completed features on the two branches are used to model the stereo consistency between the left and right views through PAM. In the following, for simplicity, we take the left view branch as an example to introduce the three main involved components.

3.2. Flow-guided Deformable Temporal Alignment

Due to the image variation caused by camera and object motion, it is difficult to directly utilize the temporal information of the reference frames to complete missing regions of the target frame. Therefore, an extra alignment module is necessary for video inpainting.

Deformable alignment has achieved a significant improvement over flow-based alignment thanks to the offset learning of the sampling convolution kernels introduced in deformable convolution (DCN). Various forms of deformable convolutional temporal alignment networks have been proposed in the past few years, such as DAPC-

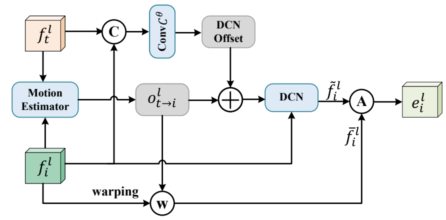

Figure 3. Illustration of the flow-guided deformable temporal alignment module.

Net [42], EDVR [40], and TDAN [36]. However, these networks often suffer from offset overflow during training, deteriorating the final alignment performance. To relieve the burden of offset learning, Chan et al. [1] used the optical flow field as base offset of deformable convolution. However, this alignment module has the following disadvantages: 1) It uses a heavyweight pre-trained neural network to generate accurate optical flow with video frames as input, which significantly increases the computational cost, and limits its practical application. In fact, as the basic offset of the deformable convolution, optical flow is more robust to errors. 2) It is achieved in an unsupervised manner, which is difficult to train. Based on this, we design a flow-guided deformable temporal alignment module to perform reference frame alignment at the feature level (Fig. 3).

Unlike the literature [1], our alignment module uses a 3-layer convolutional stack lightweight motion estimator to estimate the optical flow with features as input, which not only reduces the computational cost but also can be trained from scratch to generate more suitable optical flow for this task. In addition, we also develop an alignment loss to train the temporal alignment module in a self-supervised manner (see Section.3.5).

Specifically, for the reference frame feature $\boldsymbol{f}_i^l$ and the target feature $\boldsymbol{f}_t^l$ obtained by frame-level encoder, we first use the proposed motion estimator to calculate the optical flow $\boldsymbol{o}_{t\to i}^l$ between them, and utilize the calculated optical flow $\boldsymbol{o}_{t\to i}^l$ to warp the reference frame feature $\boldsymbol{f}_i^l$,

$$\boldsymbol{o}_{t\to i}^l = ME(\boldsymbol{f}_t^l, \boldsymbol{f}_i^l), \tag{1}$$

$$\bar{\boldsymbol{f}}_i^l = \mathcal{W}(\boldsymbol{f}_i^l, \boldsymbol{o}_{t\to i}^l), \tag{2}$$

where ME and W denote the motion estimator and warping operation, respectively. The optical flow $\boldsymbol{o}_{t\to i}^l$ are then used to compute the DCN offsets θ^l. Instead of directly computing the DCN offsets θ^l, we compute the residual of the optical flow as the DCN offsets:

$$\theta^l = \boldsymbol{o}_{t\to i}^l + \mathcal{C}^\theta(\boldsymbol{f}_i^l, \boldsymbol{f}_t^l). \tag{3}$$

Here, $\mathcal{C}^\theta$ denotes the regular convolution layer. $\theta^l = \{\triangle \boldsymbol{p}_n | n = 1, \ldots, |\mathcal{R}|\}$ denotes the offsets of the convolution kernels, where $\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (1,1)\}$ denotes a regular grid of a 3×3 kernel. Next, the aligned

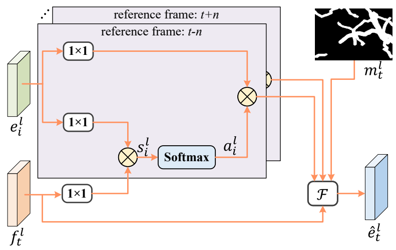

Figure 4. Illustration of the adaptive feature aggregation module.

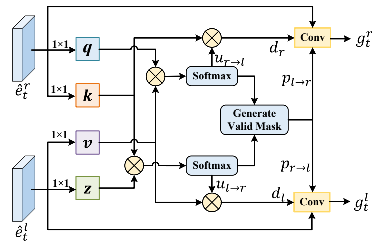

Figure 5. Illustration of the modified PAM architecture.

features $\widetilde{\boldsymbol{f}}_i^l$ of the features $\boldsymbol{f}_i^l$ can be computed by the deformable convolution:

$$\widetilde{\boldsymbol{f}}_i^l = \mathcal{DCN}(\boldsymbol{f}_i^l, \theta^l), \tag{4}$$

where $\mathcal{DCN}(\cdot)$ denotes deformable convolutional operation. Finally, to obtain more robust alignment feature, $\widetilde{\boldsymbol{f}}_i^l$ and $\overline{\boldsymbol{f}}_i^l$ are aggregated to generate the final aligned reference frame feature $\boldsymbol{e}_i^l$,

$$\boldsymbol{e}_i^l = \mathcal{A}(\widetilde{\boldsymbol{f}}_i^l, \overline{\boldsymbol{f}}_i^l), \tag{5}$$

where $\mathcal{A}$ denotes the aggregation function.

In practice, to enhance conversion flexibility and capability, we cascade two temporal alignment modules to perform feature alignment. Section.4.3 contains the ablation study on cascade operation of alignment module.

3.3. Adaptive Feature Aggregation Module

Due to occlusion, blurry regions and parallax problems, different aligned reference frames are not equally beneficial for reconstructing the missing contents in the target frame. Therefore, an adaptive feature aggregation module is used to dynamically aggregate aligned reference frames.

Specifically, as shown in Fig. 4, we first compute the similarity between each aligned reference frame feature $\boldsymbol{e}_i^l$ and the target frame feature $\boldsymbol{f}_t^l$, and then utilize *softmax* function to automatically assign aggregate weight for each aligned reference frame feature $\boldsymbol{e}_i^l$,

$$\boldsymbol{s}_i^l = \frac{exp\left((\boldsymbol{f}_t^l)^T \cdot \boldsymbol{e}_i^l\right)}{\sum_r exp\left((\boldsymbol{f}_t^l)^T \cdot \boldsymbol{e}_r^l\right)}, \tag{6}$$

where r is the number of reference frames. After obtaining the aggregated weights $\boldsymbol{s}^l$ for all aligned reference frames, the attention maps $\boldsymbol{s}_i^l$ are multiplied by the aligned reference frame feature $\boldsymbol{e}_i^l$ in a pixel-wise manner to obtain attention-modulated feature $\boldsymbol{a}_i^l$,

$$\boldsymbol{a}_i^l = \boldsymbol{s}_i^l \odot \boldsymbol{e}_i^l, \tag{7}$$

where $\odot$ denotes the element-wise multiplication. Finally, the aggregated features $\widehat{\boldsymbol{e}}_t$ are obtained by a fusion convolutional layer. Note that the missing contents in the left video may exist in the right video for the stereo video inpainting task, so it is necessary to aggregate the relevant contents in the right video to generate the missing contents in the left video. However, the direct aggregation of all right video features will increase heavy computing costs, which is not conducive to the practical application of stereo video inpainting. Therefore, when generating the missing contents of the target frame $\boldsymbol{x}_t^l$ of the left video, we only aggregate the most relevant frame $\boldsymbol{x}_t^r$ in the right video.

$$\widehat{\boldsymbol{e}}_t^l = \mathcal{F}([\boldsymbol{a}_{t-n}^l, \ldots, \boldsymbol{a}_{t+n}^l, \boldsymbol{a}_t^{r\rightarrow l}, \boldsymbol{f}_t^l, \boldsymbol{m}_t^l]), \tag{8}$$

where $\mathcal{F}$ is a 1×1 convolution layer. $\odot$ and $[\cdot,\cdot,\cdot]$ denote the element-wise multiplication and concatenation operation. $\boldsymbol{a}_t^{r\rightarrow l}$ denotes the attention-modulated features of the target frame $\boldsymbol{x}_t^r$ in the right view.

3.4. Modified PAM Architecture

In stereo image super-resolution task, Wang et al. [39] proposed the parallax attention module to estimate global matching in stereo images based on self-attention techniques [5, 47]. Since PAM can gradually focus on the features at accurate disparity using feature similarity, stereo correspondence between left and right views can then be captured. Fig. 5 depicts the structure of the redesigned PAM. For the completed feature $\widehat{\boldsymbol{e}}_t^l$ and $\widehat{\boldsymbol{e}}_t^r$ of left and right view branches, they are fed to the 1×1 convolution layer to produce the four basic elements, including $\boldsymbol{q}$, $\boldsymbol{k}$, $\boldsymbol{v}$, and $\boldsymbol{z}$. Batch-wise matrix multiplication is then performed between $\boldsymbol{q}$ and $\boldsymbol{v}$ as well as between $\boldsymbol{k}$ and $\boldsymbol{z}$, and a softmax layer is applied to generate the corresponding disparity attention maps $\boldsymbol{u}_{l\rightarrow r}$ and $\boldsymbol{u}_{r\rightarrow l}$, respectively. Next, the disparity attention maps $\boldsymbol{u}_{l\rightarrow r}$ and $\boldsymbol{u}_{r\rightarrow l}$ are respectively multiplied by $\boldsymbol{v}$ and $\boldsymbol{k}$ to produce feature $\boldsymbol{d}^l$ and $\boldsymbol{d}^r$. Note that, once $\boldsymbol{u}_{l\rightarrow r}$ and $\boldsymbol{u}_{r\rightarrow l}$ are ready, the valid masks $\boldsymbol{p}_{l\rightarrow r}$ and $\boldsymbol{p}_{r\rightarrow l}$ can be obtained by the mask generation method in reference [39]. The value of each element in the valid mask $\boldsymbol{p}_{l\rightarrow r}$ ($\boldsymbol{p}_{r\rightarrow l}$) is "0" or "1", where, "0" indicates that the pixels in the left (right) view cannot find their correspondences in the right (left) view, while "1" denotes that the pixels in the left (right) view can find their correspondences in the right (left) view. Finally, stacked feature and a valid mask are fed into a 1×1 convolution layer to generate the fused feature $\boldsymbol{g}_t^l$ and $\boldsymbol{g}_t^r$, respectively.

Table 1. Quantitative results of video inpainting on KITTI2012 and KITTI2015 datasets.

Methods	KITTI2012					KITTI2015				
	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	EPE↓	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	EPE↓
FGVC [6]	26.0814	0.8894	1.0046	0.8365	0.8832	25.8381	0.8896	0.6062	0.7296	0.8013
CPVINet [20]	26.0464	0.8729	0.8845	0.7914	0.6987	26.7131	0.8813	0.5665	0.7091	0.5502
OPN [34]	28.0218	0.9105	0.8419	0.4469	0.4586	28.7632	0.9160	0.5385	0.4092	0.3618
STTN [45]	27.6418	0.9053	0.9301	0.4750	0.4438	28.5488	0.9127	0.5942	0.4273	0.3398
FuseFormer [25]	27.4688	0.9015	0.8735	0.4090	0.4907	28.1938	0.9084	0.5667	0.5289	0.3915
E2FGVI [23]	29.3312	0.9289	0.8441	0.3557	0.5181	29.5729	0.9317	0.5407	0.3669	0.4084
FGT [48]	28.7636	0.9267	0.8073	0.3491	0.4837	29.2331	0.9304	0.5425	0.3494	0.4055
Ours	**29.6236**	**0.9303**	**0.7299**	**0.3257**	**0.3657**	**30.8191**	**0.9321**	**0.5350**	**0.2927**	**0.2668**

3.5. Loss Functions

We employ three loss functions to train the proposed network, including reconstruction loss, alignment loss, and stereo consistency loss.

Reconstruction Loss. It is used to measure pixel-level reconstruction accuracy in the whole inpainted result. In video inpainting tasks, reconstruction loss usually consists of reconstruction loss of missing regions and reconstruction loss of valid regions. The reconstruction loss of missing regions are denoted as,

$$\mathcal{L}_{hole} = \frac{\| \boldsymbol{m}_t^l \odot (\widehat{\boldsymbol{y}}_t^l - \boldsymbol{y}_t^l) \|_1}{\| \boldsymbol{m}_t^l \|_1} + \frac{\| \boldsymbol{m}_t^r \odot (\widehat{\boldsymbol{y}}_t^r - \boldsymbol{y}_t^r) \|_1}{\| \boldsymbol{m}_t^r \|_1}, \quad (9)$$

and corresponding reconstruction loss of valid regions are denoted as,

$$\begin{aligned}\mathcal{L}_{valid} = & \frac{\| (1-\boldsymbol{m}_t^l) \odot (\widehat{\boldsymbol{y}}_t^l - \boldsymbol{y}_t^l) \|_1}{\| (1-\boldsymbol{m}_t^l) \|_1} \\ & + \frac{\| (1-\boldsymbol{m}_t^r) \odot (\widehat{\boldsymbol{y}}_t^r - \boldsymbol{y}_t^r) \|_1}{\| (1-\boldsymbol{m}_t^r) \|_1},\end{aligned} \quad (10)$$

where $\odot$ indicates element-wise multiplication.

Alignment Loss. Although the proposed temporal alignment module has the potential to capture motion cues and align the reference frame and the target frame at the feature level, the implicit alignment is very difficult to learn without a supervision. To make the implicit alignment possible, we propose a self-supervised alignment loss $\mathcal{L}_{align}$ using target frame features as labels.

$$\mathcal{L}_{align} = \frac{1}{2n} \sum_{i=t-n, i\neq t}^{t+n} (\| \boldsymbol{e}_i^l - \boldsymbol{f}_t^l \|_1 + \| \boldsymbol{e}_i^r - \boldsymbol{f}_t^r \|_1), \quad (11)$$

where $\boldsymbol{e}_i^l$ and $\boldsymbol{e}_i^r$ denote the aligned reference frame feature of left and right views, respectively.

Stereo Consistency Loss. Compared with single video inpainting task, stereo video inpainting presents an additional challenge in preserving stereo consistency between left and right views. Inspired by the end-point error (EPE) [10], we propose a stereo consistency loss to measure differences between the disparity of the left and right views for ground truth and the disparity of the left and right views for the completed results. Specifically, we first calculate the optical flow $\boldsymbol{o}_y^{l\to r}$ between the left view $\boldsymbol{y}_t^l$ and the right view $\boldsymbol{y}_t^r$ of the ground truth and optical flow $\boldsymbol{o}_{\widehat{y}}^{l\to r}$ between the left view $\widehat{\boldsymbol{y}}_t^l$ and the right view $\widehat{\boldsymbol{y}}_t^r$ of the completed results. Then, the L_2-norm between $\boldsymbol{o}_y^{l\to r}$ and $\boldsymbol{o}_{\widehat{y}}^{l\to r}$ is regarded as the stereo difference between the ground truth and the completed results. The calculation formula of proposed stereo consistency loss is as follows,

$$\mathcal{L}_{stereo} = \frac{1}{H \times W \times C} \| \boldsymbol{o}_y^{l\to r} - \boldsymbol{o}_{\widehat{y}}^{l\to r} \|_2, \quad (12)$$

where $H \times W \times C$ denotes the size of the video frame $\boldsymbol{y}_t^l$.

Total Loss. The overall optimization objectives are concluded as below,

$$\mathcal{L} = \mathcal{L}_{hole} + \lambda_{valid}\mathcal{L}_{valid} + \lambda_{align}\mathcal{L}_{align} + \lambda_{stereo}\mathcal{L}_{stereo}, \quad (13)$$

where λ_{valid}, λ_{align},and λ_{stereo} are the trade-off parameters. In real implementation, we empirically set the weights of different losses as: $\lambda_{valid} = 2$, $\lambda_{align} = 0.2$, and $\lambda_{stereo} = 0.05$.

4. Experiments

4.1. Experimental Setting

Datasets. For stereo video inpainting task, there is no public dataset at present. Based on this, we designed a new stereo video inpainting (SVI) dataset using two public stereo video datasets KITTI2012 [7] and KITTI2015 [29]. Specifically, SVI includes 450 training video pairs, 135 verification video pairs and 200 test video pairs. Note that the SVI test set is divided into two parts: KITTI2012 and KITTI2015, and each part contains 100 video pairs from their original test set. The length of each video in SVI is 20 frames, which is consistent with the original KITTI dataset. As for masks, we generated two types of masks to simulate real-world applications, including stationary masks and moving masks. **Stationary masks** are used to simulate applications like watermark removal. The shapes and locations of these masks are arbitrary, and its generation process follows works [2, 48]. **moving masks** are used to simulate applications like undesired object removal and scratch restoration. Following previous single video inpainting

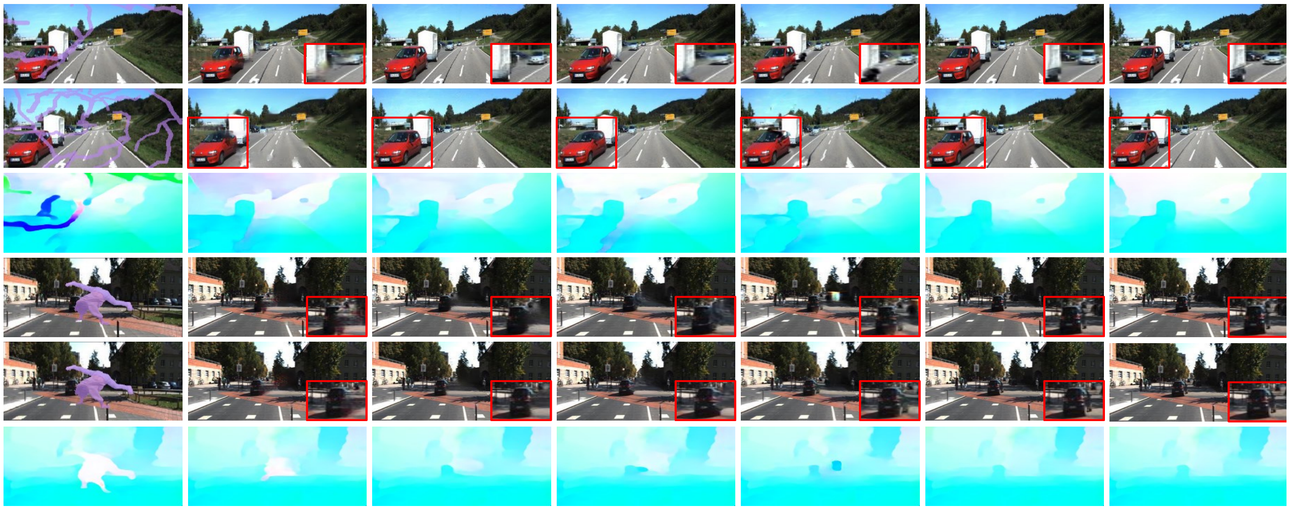

Figure 6. Qualitative results compared with single video inpainting models CPVINet [20], STTN [45], E2FGVI [23], and FGT [48]. The first and fourth lines are the left view, the second and fifth lines are the right view, and the third and sixth lines are the parallax flow between the left and right views. Better viewed at zoom level 400%.

works [21, 23, 25, 45, 51], we use the foreground object annotations in the [30] dataset as object masks, which have continuous motion and a realistic appearance. To the best of our knowledge, SVI is the first dataset for stereo video inpainting, which will be published to facilitate subsequent research and benefit other researchers.

Implementation Details. We use PyTorch to implement our model. In our experiments, an Adam optimizer with the initial learning rate of 1e-4 is used to train the proposed network, and we set $\beta_1 = 0.9$, $\beta_2 = 0.999$ as its exponential decay rates. During the training, the video sequences are resized to 256×256 as inputs. Furthermore, in our implementation, we follow the setting of signal video inpainting works [2, 17, 18, 42] to treat the $\{\boldsymbol{x}^l_{t-6}, \boldsymbol{x}^l_{t-4}, \boldsymbol{x}^l_{t-2}, \boldsymbol{x}^l_{t+2}, \boldsymbol{x}^l_{t+4}, \boldsymbol{x}^l_{t+6}\}$ as the reference frames of the target frame $\boldsymbol{x}^l_t$ in the left view. The settings in the right view are similar to those in the left view.

Baselines and Evaluation Metrics. Note that there was no work focusing on stereo video inpainting task before, so seven state-of-the-art single video inpainting methods are used as our baselines to evaluate the stereo video inpainting ability of our model, including: FGVC [6], CPVINet [20], OPN [34], STTN [45], FuseFormer [25], E2FGVI [23], and FGT [48]. To ensure the comparability of experimental results, these baselines are fine-tuned multiple by their released models and codes, and report their best results in this paper. Furthermore, we choose five metrics to report quantitative results of inpainted videos, including PSNR [8], SSIM [34], LPIPS [50], flow warping error (E_{warp}) [19], and EPE [10]. Specifically, PSNR and SSIM are frequently used metrics for distortion-oriented image and video assessment. LPIPS is a recently proposed metric to imitate human perception of image similarity. E_{warp} is employed to measure the temporal consistency. Furthermore, similar to portraying the stereo consistency in the stereo video super-resolution [12], we also compute the EPE by calculating the Euclidean distance between the disparity of the inpainted stereo frames and ground-truth frames to measure the stereo correlation of the inpainted results.

4.2. Experimental Results and Analysis

Quantitative Results. We report quantitative results of our method and other baselines on KITTI2012 [7] and KITTI2015 [29] in Tab. 1. As shown in this table, our proposed method achieves state-of-the-art results in all four evaluation metrics on two datasets compared to the single video inapinting methods. The superior results demonstrate that our method can generate videos with less distortion (PSNR and SSIM), more visually plausible contents (LPIPS), better temporal coherence (E_{warp}), and more consistent stereo correlation (EPE), which further verifies the necessity of developing stereo video inpainting model.

Qualitative Results. To further evaluate the visual quality of the inpainted stereo video, we show two examples of our model compared with four competitive single video inpainting models (including CPVINet [20], STTN [45], E2FGVI [23], and FGT [48]) in Fig. 6. As can be observed, inpainted results of the stereo video obtained by the single video inpainting model can generate specious missing contents on a single view, but fail to effectively explore the stereo cues between the left and right views. In contrast, our proposed model can not only generate vivid textures but also produce stereo consistent contents.

User Study. We conduct a user study for a more comprehensive comparison. We select three state-of-the-art single video inpainting methods as the baseline for our user study,

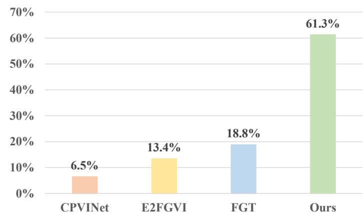

Figure 7. User preference results of four methods.

Table 2. Ablation study of alignment manner.

Index	alignment manner	PSNR↑	SSIM↑	E_warp↓	LPIPS↓	EPE↓
1	without align	25.7103	0.8786	0.6573	0.7809	0.8196
2	flow warping	29.4325	0.9277	0.5498	0.3315	0.3104
3	DCN	30.1276	0.9283	0.5474	0.3136	0.2917
4	flow + DCN	30.2413	0.9302	0.5425	0.3056	0.2798
5	ME + DCN	30.2206	0.9293	0.5432	0.3071	0.2805
6	ME + DCN + agg	30.3427	0.9308	0.5413	0.2998	0.2794
7	ME + DCN + agg + cas	30.5876	0.9315	0.5397	0.2989	0.2733
8	ME + DCN + agg + cas + $\mathcal{L}_{align}$	**30.8191**	**0.9321**	**0.5350**	**0.2927**	**0.2668**

including CPVINet [20], E2FGVI [23], and FGT [48]. 30 participants were invited to conduct a questionnaire survey for the inpainted results of 10 videos. Every volunteer is shown randomly sampled 5 video triplets and asked to select a visually better inpainting video. To ensure reliable subjective evaluation, the inpainting results obtained by the four methods are scrambled during each interrogation, and each video can be played multiple times. As shown in Fig. 7, we collected 150 votes from 30 volunteers and show the percentage of votes for each method in the form of histogram chart. The comparison results show that the proposed method can generate more visually pleasing results.

4.3. Ablation Study

Effectiveness of alignment manner. In this section, we conducted ablation research on the alignment manner of the reference frames. From Tab.2, we can obtain following conclusions: 1) The alignment module significantly improves the quality of inpainted videos; 2) The flow-guided deformable alignment manner ($4th$ and $5th$ rows) achieves superior results compared to flow-based alignment ($2th$ row) and traditional deformable alignment ($3th$ row); 3) Using the optical flow calculated by the lightweight motion estimator (ME) to guide the deformable convolution alignment will not significantly reduce the result of video inpainting ($5th$ row); 4) Aggregating $\widetilde{\boldsymbol{f}}_i^l$ and $\overline{\boldsymbol{f}}_i^l$ by Eq.5 can further improve the alignment performance of reference frames ($6th$ row); 5) The strategy of expanding the receptive field by cascading operation to improve the inpainting effect in large motion scenes can effectively ($7th$ row); 6) Using the self-supervised alignment loss $\mathcal{L}_{align}$ during training can improve the performance of the alignment module ($8th$ row).

Effectiveness of PAM. As described in Section.3.4, PAM is used to model the stereo correspondence between the left and right views. In this section, we investigate the effectiveness of this module in the stereo video inpainting task

Table 3. Effectiveness of PAM and $\mathcal{L}_{stereo}$.

	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	EPE↓
CPVINet	26.7131	0.8813	0.5665	0.7091	0.5502
CPVINet+PAM	26.9352	0.8896	0.5580	0.6914	0.5399
CPVINet+PAM+$\mathcal{L}_{stereo}$	27.0583	0.8909	0.5501	0.6877	0.5273
OPN	28.7632	0.9160	0.5385	0.4092	0.3618
OPN+PAM	28.9103	0.9202	0.5269	0.4003	0.3489
OPN+PAM+$\mathcal{L}_{stereo}$	29.0562	0.9288	0.5205	0.3913	0.3235
STTN	28.5488	0.9127	0.5942	0.4273	0.3398
STTN+PAM	28.7209	0.9196	0.5817	0.4139	0.3209
STTN+PAM+$\mathcal{L}_{stereo}$	28.9897	0.9205	0.5786	0.4057	0.3077
w/o $\mathcal{L}_{stereo}$	30.5691	0.9306	0.5394	0.3019	0.3065
Full model	**30.8191**	**0.9321**	**0.5350**	**0.2927**	**0.2668**

Table 4. Ablation study of cross view aggregation strategy.

Index	Method	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	EPE↓
1	w/o across views	30.6875	0.9316	0.5385	0.3006	0.2714
2	Full model	**30.8191**	**0.9321**	**0.5350**	**0.2927**	**0.2668**

by adding PAM to the single video inpainting network. As shown in Tab. 3, compared with the original single video inpainting model, the stereo video inpainting performance of the model with PAM is improved, especially the stereo correlation (EPE) between the left and right views.

Effectiveness of $\mathcal{L}_{stereo}$. $\mathcal{L}_{stereo}$ is used to regularize the trained parameters, so that the trained model is able to yield high-quality stereo video inpainting results with better stereo consistency. In Tab. 3, we conducted an ablation study on $\mathcal{L}_{stereo}$. As we can see, models with $\mathcal{L}_{stereo}$ participating in training can obtain smaller EPE indicators. This shows that the designed $\mathcal{L}_{stereo}$ is effective in preserving the stereo consistency of stereo video inpainting results.

Necessity of aggregate across views. As mentioned in Section.3.3, we used the relevant information from the right view when generating the missing contents of the left view branch. Tab. 4 studies the effectiveness of this cross view aggregation strategy. From Tab. 4, we can observe that the model using cross view aggregation strategy has better inpainting results. This indicates that it is necessary to aggregate information across views in stereo video inpainting.

5. Conclusion

In this work, we studied stereo video inpainting, attempting to inpaint the missing regions of the left and right video, while maintaining their temporal and stereo consistency. To achieve this, we propose a novel deep network architecture for stereo video inpainting, named SVINet. SVINet first generates missing contents on the left and right view branches through the classic "alignment–aggregation" pipeline. Then the completed results of the left and right view branches are fed into the PAM to model the stereo correlation between views. Furthermore, we also design a stereo consistency loss to regularize the trained parameters, so that our model is able to yield high-quality stereo video inpainting results with better stereo consistency. Experimental results show that the proposed method is effective in stereo video inpainting.

References

[1] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5962–5971, 2022. 4

[2] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9066–9075, 2019. 1, 2, 6, 7

[3] Canqiang Chen, Chunmei Qing, Xiangmin Xu, and Patrick Dickinson. Cross parallax attention network for stereo image super-resolution. *IEEE Transactions on Multimedia*, 24:202–216, 2022. 2

[4] Shen Chen, Wei Ma, and Yue Qin. Cnn-based stereoscopic image inpainting. In Yao Zhao, Nick Barnes, Baoquan Chen, Rüdiger Westermann, Xiangwei Kong, and Chunyu Lin, editors, *Image and Graphics*, pages 95–106, 2019. 3

[5] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3141–3149, 2019. 5

[6] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 713–729, 2020. 1, 2, 3, 6, 7

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 6, 7

[8] Zhang Haotian, Mai Long, Wang Hailin, JinZha ando wen, and Ning Xu; John Collomosse. An internal learning approach to video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2720–2729, 2019. 7

[9] Alexandre Hervieu, Nicolas Papadakis, Aurelie Bugeau, Pau Gargallo, and Vicent Caselles. Stereoscopic image inpainting: Distinct depth maps and images inpainting. In *2010 20th International Conference on Pattern Recognition*, pages 4101–4104, 2010. 3

[10] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. 2, 6, 7

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[12] Hassan Imani, Md Baharul Islam, and Lai-Kuan Wong. A new dataset and transformer for stereoscopic video super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 705–714, 2022. 7

[13] Zhong Ji, Jiacheng Hou, Yimu Su, Yanwei Pang, and Xuelong Li. G2lp-net: Global to local progressive video inpainting network. *IEEE Trans. Circuits Syst. Video Technol.*, page Early Access, 2022. 1

[14] Jaeyeon Kang, Seoung Wug Oh, and Seon Joo Kim. Error compensation framework for flow-guided video inpainting. 2022. 2

[15] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Occlusion-aware video object inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[16] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep blind video decaptioning by temporal aggregation and recurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4263–4272, 2019. 2

[17] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Recurrent temporal aggregation framework for deep video inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(5):1038–1052, 2020. 7

[18] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5792–5801, 2019. 1, 7

[19] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 179–195, 2018. 7

[20] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4413–4421, 2019. 1, 3, 6, 7, 8

[21] Ang Li, Shanshan Zhao, Xingjun Ma, M. Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and R. Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 7

[22] Ang Li, Shanshan Zhao, Qingjie Zhang, and Qiuhong Ke. Iterative geometry-aware cross guidance network for stereo image inpainting. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 2, 3

[23] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17562–17571, 2022. 1, 2, 6, 7, 8

[24] Miao Liao, Feixiang Lu, Dingfu Zhou, Sibo Zhang, Wei Li, and Ruigang Yang. Dvi: Depth guided video inpainting for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–17, 2020. 1

[25] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 14040–14049, 2021. 3, 6, 7

[26] Ruixin Liu, Bairong Li, and Yuesheng Zhu. Temporal group fusion network for deep video inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3539–3551, 2022. 2

[27] Ruixin Liu, Zhenyu Weng, Yuesheng Zhu, and Bairong Li. Temporal adaptive alignment network for deep video inpainting. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 927–933, 2020. 3

[28] Wei Ma, Mana Zheng, Wenguang Ma, Shibiao Xu, and Xiaopeng Zhang. Learning across views for stereo image completion. *IET Computer Vision*, 14(7):482–492, 2020. 3

[29] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 6, 7

[30] F. Perazzi, J. Pont-Tuset, B. Mcwilliams, L. Van Gool, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016. 7

[31] Félix Raimbault and Anil Kokaram. Stereo video inpainting. In *Stereoscopic Displays and Applications XXII*, volume 7863, pages 426–438. SPIE, 2011. 1, 3

[32] Félix Raimbault, François Pitié, and Anil Kokaram. Stereo video completion for rig and artefact removal. In *2012 13th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 1–4, 2012. 1, 3

[33] Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3511–3520, 2022. 3

[34] Oh Seoung, Wug, Lee Sungho, Lee Joon-Young, and Kim Seon, Joo. Onion-peel networks for deep video completion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4402–4411, 2019. 1, 3, 6, 7

[35] Vishnu Sanjay Ramiya Srinivasan, Rui Ma, Qiang Tang, Zili Yi, and Zhan Xu. Spatial-temporal residual aggregation for high resolution video inpainting. *arXiv preprint arXiv:2111.03574*, 2021. 3

[36] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3369, 2020. 4

[37] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5232–5239, 2019. 2

[38] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 3

[39] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12251, 2019. 2, 5

[40] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 4

[41] Zhiliang Wu, Changchang Sun, Hanyu Xuan, Kang Zhang, and Yan Yan. Divide-and-conquer completion network for video inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, Early Access, 2022. 3

[42] Zhiliang Wu, Kang Zhang, Hanyu Xuan, Jian Yang, and Yan Yan. Dapc-net: Deformable alignment and pyramid context completion networks for video inpainting. *IEEE Signal Processing Letters*, 28:1145–1149, 2021. 1, 4, 7

[43] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3723–3732, 2019. 2, 3

[44] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020. 2

[45] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3723–3732, 2020. 3, 6, 7

[46] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1494, 2019. 1

[47] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363, 2019. 5

[48] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3, 6, 7, 8

[49] Kaidong Zhang, Jingjing Fu, and Dong Liu. Inertia-guided flow completion and style fusion for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5982–5991, 2022. 2

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7

[51] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 7