DLVINet: Advancing Dual-Lens Video Inpainting Beyond Parallax Constraints

Zhiliang Wu[1], Kun Li[2], Yunqiu Xu[1], Hehe Fan[1], Yi Yang[1]*

[1] CCAI, Zhejiang University, China
[2] Department of Computer Science, Hong Kong Baptist University, China

Abstract

Dual-lens video inpainting aims to simultaneously restore missing or corrupted contents in videos captured by each lens of binocular systems. Although preliminary explorations have been conducted, existing methods still face two key challenges: limited exploitation of long-range reference information and inadequate modeling of inter-lens consistency in non-standard binocular systems. In this paper, we propose a novel dual-lens video inpainting framework named DLVINet, which addresses these challenges with two core components. Firstly, we develop a sparse spatial-temporal transformer (SSTT) that effectively utilizes the information from distant frames to complete the video contents of each lens individually. By employing sparse spatial-temporal attention with a channel selection mechanism, SSTT not only restores missing regions, but also avoids introducing redundant or irrelevant information. Furthermore, SSTT introduces a multi-scale feed-forward network to enrich the multi-scale representation of completed features. Secondly, we design a cross-lens texture transformer (CLTT) to model inter-lens consistency. By interacting with corresponding features between lenses under the guidance of cross-attention, CLTT captures global inter-lens correspondences. Such a design enables effective cross-view information modeling without being constrained by horizontal parallax, which is particularly critical for non-standard binocular systems. Extensive experiments demonstrate the effectiveness of our DLVINet.

1 Introduction

Video inpainting is a fundamental task in computer vision (Xu et al. 2022b; Li et al. 2025c; Xu et al. 2022a; Quan et al. 2021b; Li et al. 2024b; Xu et al. 2025), which aims to fill in missing or corrupted regions in a video with plausible contents. Although existing methods (Hou et al. 2024; Wu et al. 2023b; Bian et al. 2025) have achieved significant progress, they usually focus on single-lens video inpainting task. With the widespread use of smartphones, unmanned aerial vehicles and autonomous robots, dual-lens vision processing techniquehas attracted increasing attention from researchers, including dual-lens video inpainting. Compared to single-lens video inpainting, dual-lens video inpainting also needs to consider the inter-lens consistency of inpainted contents. Naively applying these single-lens video inpainting algorithm to fill in missing or corrupted regions of dual-lens video will cause serious inter-lens inconsistency. Therefore, it is necessary and promising to explore inpainting techniques applicable to dual-lens videos.

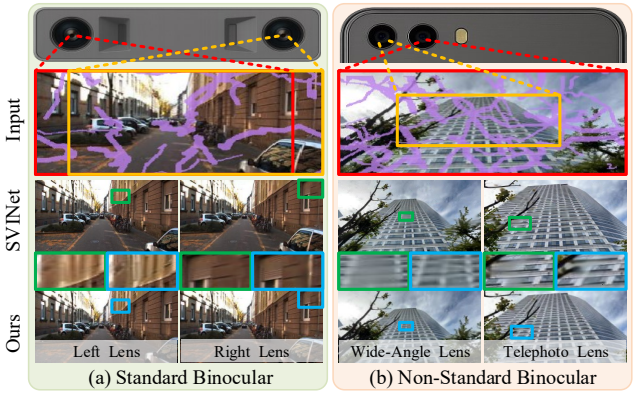

Figure 1: Examples of dual-lens video inpainting under two scenarios: (a) standard binocular system and (b) non-standard binocular system. As shown here, SVINet (Wu et al. 2023a) generates blurry inpainted contents and fails to maintain the inter-lens consistency under non-standard binocular systems. Conversely, our method not only produces texture details, but also effectively models inter-lens consistency in both standard and non-standard scenarios.

Recently, Wu et al. (Wu et al. 2023a) proposed the first dual-lens video inpainting network named SVINet suitable for stereo video scenarios. SVINet first employed the "alignment–aggregation" technical pipeline to restore the intra-lens missing contents, and then used a modified parallax attention module (PAM) (Wang et al. 2022) to model the inter-lens consistency. Despite its impressive performance, SVINet fails to make effective use of long-term reference information in videos. In fact, the "alignment–aggregation" pipeline typically restricts the range of reference frames to the nearby (short-term) frames of the target frame, which prevents available contents on distant frames from being propagated into the missing regions of target frame. This limitation is critical when dealing with scenes containing large or slow-moving objects, since short-term frames alone

*Corresponding author.

cannot provide sufficient reference information to restore the missing regions (Zhang et al. 2024a; Wang et al. 2023).

Additionally, SVINet lacks the ability to maintain inter-lens consistency when dealing with video captured by non-standard binocular systems. Essentially, the modified PAM in SVINet usually focuses on capturing consistency between lenses with limited horizontal parallax (Zhang et al. 2024b). Such a design is only suitable for videos captured by standard binocular systems with the same field of view (FoV) and optical zoom factor (Chen et al. 2022), such as the stereo video illustrated in Fig. 1(a). However, in real-world applications, non-standard binocular systems are more common, such as dual-camera configurations on mobile phones that combine wide-angle and telephoto lenses. As shown in Fig. 1(b), videos captured by such systems typically have different FoV and optical zoom factor, resulting in significant parallax in both horizontal and vertical directions. In such scenarios, it is difficult for the modified PAM to effectively capture inter-lens consistency. This limitation restricts the applicability of SVINet in scenarios involving non-standard binocular systems.

In this paper, we propose a novel framework for dual-lens video inpainting, called DLVINet. Our DLVINet consists of a sparse spatial-temporal transformer (SSTT) and a cross-lens texture transformer (CLTT), aiming to address the aforementioned challenges from the following two aspects:

(1) We design a SSTT that makes full use of the reference information of distant frames to generate missing or corrupted contents. Unlike existing spatial-temporal transformers (Zeng et al. 2020; Liu et al. 2021; Li et al. 2022b; Zhang et al. 2024a; Zhou et al. 2023) that aggregate the features by all attention relations based on query-key pairs, our SSTT leverages sparse spatial-temporal attention with a channel selection mechanism to aggregate only the most relevant channels into missing or corrupted regions. Such a strategy not only completes the global structure of the missing or corrupted regions, but also avoids introducing redundant or irrelevant contents. In addition, we introduce a multi-scale feed-forward network to further explore and enhance the multi-scale information in the completed features, thereby aggregating richer multi-scale cues that are crucial for latent frame completion into the global representation.

(2) We develop a CLTT to better model the inter-lens consistency of video pairs. Unlike PAM (Wang et al. 2022) and modified PAM (Wu et al. 2023a), which focus solely on pixel-to-pixel correspondences, our CLTT learns global inter-lens correspondences by interacting with corresponding features between two lenses under the guidance of cross-attention. In this way, CLTT adaptively adjusts receptive fields across different parallax ranges by selecting larger weights based on disparity shifts obtained from cross-attention, thereby enabling effective cross-view information integration without being constrained by horizontal parallax. This design significantly enhances the generality of our framework in real-world applications, especially when processing videos captured by non-standard binocular systems, as illustrated in Fig. 1 (b).

Experimental results show that our DLVINet achieves SOTA inpainting results on videos captured by both standard and non-standard binocular systems. Compared with the best dual-lens video inpainting baseline, our DLVINet improves the inter-lens consistency by 14.98% on standard binocular videos and 30.37% under non-standard settings.

To sum up, our contributions are summarized as follows:

- We propose a novel framework for dual-lens video inpainting, which handles videos captured by both standard and non-standard binocular systems. To the best of our knowledge, this is the first dual-lens video inpainting framework involving non-standard binocular systems.
- We develop a sparse spatial-temporal transformer and a cross-lens texture transformer. The former aims to effectively utilize reference information from distant frames, while the latter focuses on modeling inter-lens consistency without being constrained by horizontal parallax.
- Extensive experiments demonstrate the superiority of our method in the dual-lens video inpainting task. Notably, our method also sheds light on the subsequent research of general dual-lens video inpainting.

2 Related Works

Single-Lens Video Inpainting. Benefiting from deep learning, several single-lens video inpainting methods have achieved great progress.These methods can be divided into two lines: flow-based methods and pixel-oriented methods.

Flow-based methods (Gao et al. 2020; Li et al. 2022b; Xu et al. 2019; Zou et al. 2021; Cho et al. 2025) first utilize a deep flow completion network to restore the flow between frames. Subsequently, they employ the restored flow to guide corresponding pixels from neighboring frames towards the missing regions, thereby generating the missing contents. However, due to their limited temporal receptive field, these methods struggle to capture information from distant frames. Consequently, their performance degrades in scenes with large objects or slow motion (Quan et al. 2024).

The second category focuses on directly synthesizing the missing contents in video frames by employing various types of spatial-temporal context aggregation. For instance, Wang et al. (Wang et al. 2019), Chang et al. (Chang et al. 2019) and Kim et al. (Kim et al. 2020) aggregated the temporal information in a local temporal window using 3D CNNs to generate missing contents of the video frames. Furthermore, some works (Wu et al. 2025b; Cai et al. 2022; Wu et al. 2023c; Liu et al. 2021; Wu et al. 2024a) incorporated attention mechanisms (Li et al. 2025b; Chen et al. 2025) into CNN-based networks to extend the limited receptive field of the temporal domain. Among these methods, Zeng et al. (Zeng et al. 2020), Liu et al. (Liu et al. 2021), Zhou et al. (Zhou et al. 2023), Zhang et al. (Zhang et al. 2024a), and Wu et al. (Wu et al. 2024b) employed transformers (Hua et al. 2025; Fan et al. 2021; Li et al. 2024a; Liang et al. 2023b) to retrieve similar features in a considerable temporal receptive field, resulting in high-quality single-lens video inpainting. Recently, several researchers (Zhang et al. 2024c; Lee et al. 2025; Sun et al. 2025; Yang et al. 2025; Zi et al. 2025) have explored the application of diffusion models in video inpainting. Despite the unprecedented performance achieved by these methods, directly extending them

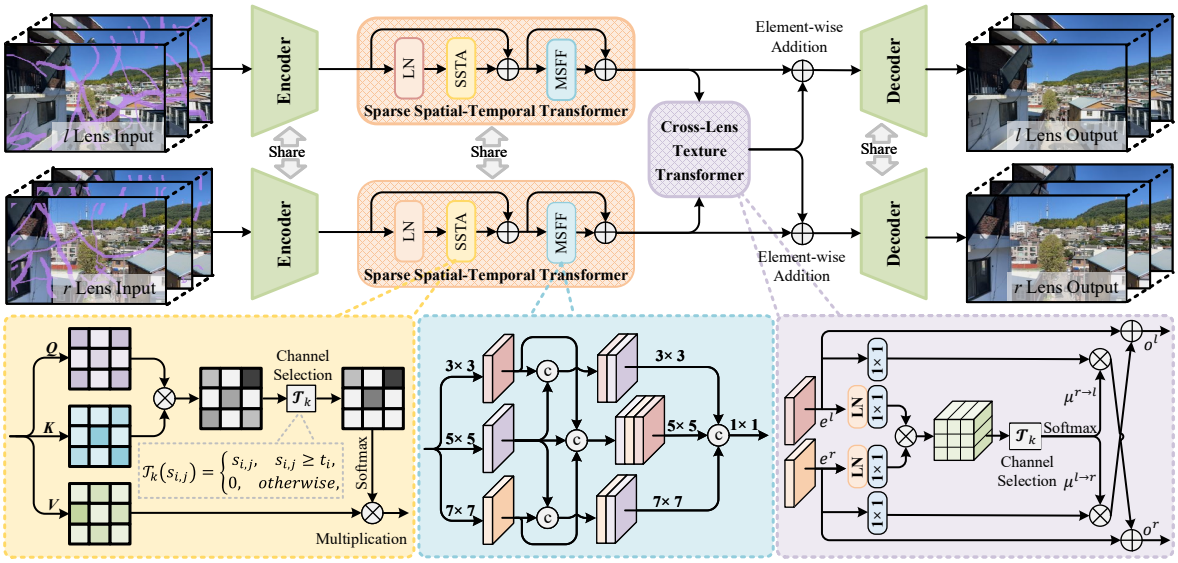

Figure 2: Illustration of the proposed dual-lens inpainting network (DLVINet). The core components of our DLVINet contains a sparse spatial-temporal transformer (**SSTT**) with sparse spatial-temporal attention (SSTA) and multi-scale feed-forward network (MSFF), and a cross-lens texture transformer (**CLTT**).

to dual-lens video inpainting will lead to inconsistent correspondence between lenses.

Dual-Lens Image/Video Inpainting. Dual-lens image inpainting is a sub-task of image inpainting, and several methods (Chen et al. 2019; Ma et al. 2020; Li et al. 2022a) have been proposed. However, naively using them to fill in missing regions of dual-lens video in a frame-by-frame manner will lose inter-frame motion continuity, resulting in flicker artifacts (Zhang et al. 2023b; Wang et al. 2024). Recently, Wu et al. (Wu et al. 2023a) developed a dual-lens video inpainting network named SVINet, which can maintain inter-lens consistency while ensuring the temporal consistency by an attention mechanism (Quan et al. 2021a; Liu et al. 2025). However, SVINet is only suitable for dual-lens videos captured by standard binocular systems, fails to model the inter-lens consistency in non-standard setting.

Top-k selection. Zhao et al. (Zhao et al. 2020) first attempt of the top-k selection was used in the NLP task. soon afterwards, researchers further introduced k-NN attention in various computer vision tasks to boost vision Transformers (Liang et al. 2022; Fan et al. 2022; Liang et al. 2023a; Wang et al. 2025), such as image restoration (Zhang et al. 2023a), and video super-resolution (Tuo et al. 2023). Unlike these methods that perform top-k selection in the *spatial dimension*, we design an efficient top-k *channel selection* operator. Such a design helps to simultaneously complete the global semantics and local details of missing regions in a sparse manner, which is crucial for video inpainting.

3 Proposed Method

3.1 Formulation and Overview

Assume $(\boldsymbol{X}^l, \boldsymbol{X}^r) = \{(\boldsymbol{x}_1^l, \boldsymbol{x}_1^r), (\boldsymbol{x}_2^l, \boldsymbol{x}_2^r), \ldots, (\boldsymbol{x}_T^l, \boldsymbol{x}_T^r)\}$ is the corrupted dual-lens video with sequence length T, where $\boldsymbol{x}_i^l$ and $\boldsymbol{x}_i^r$ denote the i-th corrupted frame of first lens video $\boldsymbol{X}^l$ and second lens video $\boldsymbol{X}^r$, respectively. The corresponding frame-wise missing or corrupted regions are denoted by the binary mask sequence $(\boldsymbol{M}^l, \boldsymbol{M}^r) = \{(\boldsymbol{m}_1^l, \boldsymbol{m}_1^r), (\boldsymbol{m}_2^l, \boldsymbol{m}_2^r), \ldots, (\boldsymbol{m}_T^l, \boldsymbol{m}_T^r)\}$. For each binary mask $\boldsymbol{m}_i$, "1" indicates the missing or corrupted pixel, and "0" represents the valid (uncorrupted) pixel. The goal of dual-lens video inpainting is to generate an completed dual-lens video sequence $(\widehat{\boldsymbol{Y}}^l, \widehat{\boldsymbol{Y}}^r) = \{(\widehat{\boldsymbol{y}}_1^l, \widehat{\boldsymbol{y}}_1^r), (\widehat{\boldsymbol{y}}_2^l, \widehat{\boldsymbol{y}}_2^r), \ldots, (\widehat{\boldsymbol{y}}_T^l, \widehat{\boldsymbol{y}}_T^r)\}$, which should be spatially, temporally, and inter-lens consistent with original video sequence $(\boldsymbol{Y}^l, \boldsymbol{Y}^r) = \{(\boldsymbol{y}_1^l, \boldsymbol{y}_1^r), (\boldsymbol{y}_2^l, \boldsymbol{y}_2^r), \ldots, (\boldsymbol{y}_T^l, \boldsymbol{y}_T^r)\}$.

As illustrated in Fig. 2 (a), the proposed dual-lens video inpainting network (DLVINet) mainly consists of four parts: frame-level encoder, sparse spatial-temporal transformer (SSTT), cross-lens texture transformer (CLTT), and frame-level decoder. SSTT and CLTT are the core components of our DLVINet. The former learns a joint sparse spatial-temporal transformation to complete missing regions in each lens sequence, while the latter facilitates interaction between the two lenses and models inter-lens consistency.

3.2 Sparse Spatial-Temporal Transformer

Transformer (Gu et al. 2025; Li et al. 2025a) has garnered increasing attention from researchers in the field of video inpainting due to its significant advantages in long-range modeling capacity. Although transformer-based video inpainting methods (Li et al. 2022b; Zhou et al. 2023; Zhang et al. 2024a; Ji et al. 2024; Wu et al. 2024b) have shown promising results, they still suffer from two major limitations. ***First***, these methods aggregate features using all attention relations based on query-key pairs to generate the missing contents. Such aggregation often leads to redundant contents being

filled into the missing regions, resulting in suboptimal results (Li et al. 2023). ***Second***, these studies typically use the single-scale depth-wise convolutions in the feed-forward process to update features of corrupted regions. This operation tends to ignore correlations between missing contents across scales. In fact, rich multi-scale information has been demonstrated to significantly improve the performance of vision restoration tasks (Liu et al. 2024b; Zhou et al. 2024).

To this end, we propose a sparse spatial-temporal transformer (SSTT) to complete the video content of each lens individually. Unlike above standard transformer, the proposed SSTT first employs a sparse spatial-temporal attention (SSTA) module, which utilizes a top-k channel selection strategy to aggregate the most relevant channels into the missing regions. The resulting features are then refined through a multi-scale feed-forward (MSFF) network to update the embeddings of the corrupted regions. This design not only avoids introducing redundant content into the missing regions, but also enhances the multi-scale representation of the corrupted embeddings. In the following, we take the $\boldsymbol{X}^l$ branch as an example to introduce SSTA and MSFF.

Sparse Spatial-Temporal Attention (SSTA). Let $\boldsymbol{F}^l = \{\boldsymbol{f}_1^l, \boldsymbol{f}_2^l, \cdots, \boldsymbol{f}_T^l\}$ denotes the features encoded by the frame-level encoder for $\boldsymbol{X}^l$, where $\boldsymbol{f}_i^l \in \mathbb{R}^{h \times w \times c}$. We first project the features $\boldsymbol{f}_i^l$ into $\boldsymbol{Q}_i$(query), $\boldsymbol{K}_i$(key), and $\boldsymbol{V}_i$(value),

$$\boldsymbol{Q}_i, (\boldsymbol{K}_i, \boldsymbol{V}_i) = \mathcal{G}_q(\boldsymbol{f}_i^l), (\mathcal{G}_k(\boldsymbol{f}_i^l), \mathcal{G}_v(\boldsymbol{f}_i^l)), \tag{1}$$

where $1 \leq i \leq T$. $\mathcal{G}_q(\cdot)$, $\mathcal{G}_k(\cdot)$ and $\mathcal{G}_v(\cdot)$ denote the 1×1 2D convolution. After obtaining three basic elements, we extract the spatial patches with shape $r_1 \times r_2 \times c$ from $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$ and $\boldsymbol{V}_i$ to calculate the similarity matrix $\boldsymbol{S}$ between query-key pairs. The similarities between i-th query patch ($\boldsymbol{q}_i$) and j-th key patch ($\boldsymbol{k}_j$) can be calculated by matrix multiplication,

$$s_{i,j} = \frac{\boldsymbol{q}_i \cdot (\boldsymbol{k}_j)^T}{\sqrt{r_1 \times r_2 \times c}}, \tag{2}$$

where $1 \leq i, j \leq N$, $N = T \times \frac{h}{r_1} \times \frac{w}{r_2}$, and $s_{i,j}$ is the element of similarity matrix $\boldsymbol{S}$ at position (i, j).

Then, a top-k selection operator $\mathcal{T}_k(\cdot)$ is applied along the *channel dimension* to preserve the most significant contents and remove the useless ones. Unlike the Dropout strategy that randomly deactivates a subset of neurons, our $\mathcal{T}_k(\cdot)$ selectively retains the most significant components based on the similarity matrix $\boldsymbol{S}$, aiming to remove the useless ones (Chen et al. 2021; Li et al. 2023). This dynamic selection makes the similarity matrix $\boldsymbol{S}$ from *dense* to *sparse*.

$$\alpha_{i,j} = \mathcal{T}_k(s_{i,j}) = \begin{cases} s_{i,j}, & s_{i,j} \geq t_i, \\ 0, & otherwise, \end{cases} \tag{3}$$

where t_i is the k-th largest value in the j-th row of the similarity matrix $\boldsymbol{S}$. k is an adjustable parameter to dynamically control the magnitude of sparsity. The attention weights of all patches can be calculated by a *softmax* function:

$$\gamma_{i,j} = \begin{cases} \exp(\alpha_{i,j}) / \sum_{t=1}^{N} \exp(\alpha_{i,t}), & \boldsymbol{q}_i \in \Omega, \\ 0, & otherwise, \end{cases} \tag{4}$$

where Ω denotes the uncorrupted regions. In fact, only uncorrupted contents is useful when completing the missing regions. Finally, the query of each patch can be obtained by using an attention-weighted summation of the values:

$$\widehat{\boldsymbol{v}}_i = \sum_{j=1}^{N} \gamma_{i,j} \boldsymbol{v}_j, \tag{5}$$

where $\boldsymbol{v}_j$ denotes the value of the j-th patch. The feature $\widehat{\boldsymbol{V}}_i$ is acquired by piecing all $\widehat{\boldsymbol{v}}_i$ together.

Multi-Scale Feed-Forward (MSFF). After obtaining the feature $\widehat{\boldsymbol{V}}_i$, we feed it into a MSFF network to update the features of the corrupted regions, enabling the model to leverage multi-scale information for improved inpainting results. As shown in Fig. 2 (d), we first expand the channel dimension of the features $\widehat{\boldsymbol{V}}_i \in \mathbb{R}^{h \times w \times c}$ from c to $3c$ using a 1×1 convolution layer. Then, the acquired features is split into three tensors along the channel dimension and fed into three parallel branches, each with an input size of $h \times w \times c$. In three parallel branches, we employ 3×3, 5×5 and 7×7 depth-wise convolutions to capture local information at three scales, respectively. During the feed-forward process, features from different branches interact with one another, thereby enhancing multi-scale information extraction. Finally, we concatenated the output of three branches and shrink the channels via another 1×1 convolution layer to match the dimension of the input channels. The entire feed-forward process can be defined as:

$$\boldsymbol{e}_i = \mathcal{MSFF}(\widehat{\boldsymbol{V}}_i). \tag{6}$$

3.3 Cross-Lens Texture Transformer

Directly decoding the feature $\boldsymbol{e}_i$ obtained from Eq. (6) to generate the final results often leads to severe inter-lens inconsistency, which is crucially detrimental to the dual-lens video inpainting task. Therefore, it is both necessary and urgent to explore the relevant information between lenses to ensure inter-lens consistency of the inpainted results. In fact, inter-lens consistency modeling has been studied preliminary in the stereo image super-resolution task. For instance, Wang et al. (Wang et al. 2022), Cheng et al. (Cheng et al. 2023), and Liu et al. (Liu et al. 2024a) have designed the parallax attention, cross-stereo attention, and cascaded parallax attention module to model inter-lens consistency. Furthermore, SVINet (Wu et al. 2023a) developed a modified parallax attention module to model inter-lens consistency for dual-lens video inpainting task. However, these modules usually focus on modeling consistency between two-lens views with limited horizontal parallax (Zhang et al. 2024b). As a result, they are only suitable for the video captured by standard binocular systems with the same FoV and optical zoom factor, fail to model correspondence in videos obtained from non-standard binocular systems with different FoV and optical zoom factor.

To mitigate the above challenge, we propose a cross-lens texture transformer (CLTT).Unlike existing methods that focus solely on exploring pixel-to-pixel correspondences, CLTT leverages cross-attention to learn global correspondence by interacting with corresponding features between

Method		Standard Binocular Systems					Non-Standard Binocular Systems				
		PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	EPE↓	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	EPE↓
FGVC	[ECCV2020]	26.0814	0.8894	1.0046	0.8365	0.8832	26.9380	0.8848	0.1463	0.6919	9.7888
CPVINet	[ICCV2019]	26.0464	0.8729	0.8845	0.7914	0.6987	28.1522	0.8992	0.1416	0.6002	8.1261
OPN	[ICCV2019]	28.0218	0.9105	0.8419	0.4469	0.4586	30.1130	0.9273	0.1123	0.3447	6.0247
STTN	[ECCV2020]	27.6418	0.9053	0.9301	0.4750	0.4438	30.6486	0.9292	0.1116	0.4325	6.7480
FuseFormer	[ICCV2021]	27.4688	0.9015	0.8735	0.4090	0.4907	28.7713	0.8969	0.1366	0.4848	7.5622
E2FGVI	[CVPR2022]	29.3312	0.9289	0.8441	0.3557	0.5181	29.6819	0.9174	0.1261	0.4268	6.8690
ProPainter	[ICCV2023]	28.0898	0.9043	0.8426	0.4192	0.5147	29.5802	0.9231	0.1285	0.4284	6.2707
WaveFormer	[AAAI2024]	28.2999	0.9127	0.8372	0.3776	0.5343	30.1292	0.9181	0.0959	0.3681	6.5255
DiffuEraser	[arXiv2025]	30.0193	0.9312	0.7207	0.3068	0.5214	30.9278	0.9301	0.0957	0.3591	6.5927
VideoPainter	[arXiv2025]	29.9342	0.9289	0.7196	0.3101	0.5120	30.8960	0.9299	0.0963	0.3604	6.6149
SVINet	[CVPR2023]	29.6236	0.9303	0.7299	0.3257	0.3657	29.9873	0.9232	0.1313	0.5404	8.2918
Ours		**32.2054**	**0.9618**	**0.7039**	**0.2929**	**0.3109**	**31.6228**	**0.9438**	**0.0920**	**0.3436**	**5.7732**

Table 1: Quantitative results under standard binocular systems and non-standard binocular systems.

two lenses. This design enables adaptive adjustment of receptive fields across different parallax ranges, thereby facilitating effective cross-view information integration and robust inter-lens consistency modeling without being constrained by horizontal parallax.

As shown in Fig. 2 (e), given the $\boldsymbol{e}_i^l$ and $\boldsymbol{e}_i^r$ obtained by Eq. (6) on first and second lens branches, we first calculate the relevance $\boldsymbol{a}_i$ between the two:

$$\boldsymbol{a}_i = \mathcal{H}_q(\mathcal{LN}(\boldsymbol{e}_i^l)) \cdot \mathcal{H}_k(\mathcal{LN}(\boldsymbol{e}_i^r))^T / \sqrt{c}, \tag{7}$$

where $\mathcal{LN}(\cdot)$ is the layer normalization, $\mathcal{H}_q(\cdot)$ and $\mathcal{H}_k(\cdot)$ denote the 1×1 2D convolution, and c represents the number of channels in the feature maps. Then, the bidirectional cross-attention maps between first and second lens can be calculated by a $softmax(\cdot)$ function,

$$\begin{aligned} \boldsymbol{u}_i^{r\to l} &= softmax(\mathcal{T}_k(\boldsymbol{a}_i)), \\ \boldsymbol{u}_i^{l\to r} &= softmax(\mathcal{T}_k(\boldsymbol{a}_i^T)), \end{aligned} \tag{8}$$

where $\mathcal{T}_k(\cdot)$ denotes the top-k selection adaptive operator and T is the transpose operation for matrices. Here, we exploit the same top-k selection strategy as the SSTT to adaptively select information for across-lens interactions. This strategy helps to suppress the useless information during cross-lens interaction.

The calculated map $\boldsymbol{u}_i^{r\to l}$ and $\boldsymbol{u}_i^{l\to r}$ are multiplied by their corresponding features to generate cross-lens information. Finally, the obtained cross-lens information and intra-lens information are fused to generate the final output:

$$\begin{aligned} \boldsymbol{o}_i^l &= \beta_l \boldsymbol{u}_i^{r\to l} \cdot \mathcal{H}_v^r(\boldsymbol{e}_i^l) + \boldsymbol{e}_i^l, \\ \boldsymbol{o}_i^r &= \beta_r \boldsymbol{u}_i^{l\to r} \cdot \mathcal{H}_v^l(\boldsymbol{e}_i^r) + \boldsymbol{e}_i^r, \end{aligned} \tag{9}$$

where β_l and β_r are trainable channel-wise scale parameters. $\mathcal{H}_v^l(\cdot)$ and $\mathcal{H}_v^r(\cdot)$ denote the 1×1 2D convolution that project the features $\boldsymbol{e}_i^r$ and $\boldsymbol{e}_i^l$ into value vector. The inpainted frames $\widehat{\boldsymbol{y}}_i^l$ and $\widehat{\boldsymbol{y}}_i^r$ can be generated by decoding $\boldsymbol{o}_i^l$ and $\boldsymbol{o}_i^r$ with a shared frame-level decoder.

4 Experiments

4.1 Experimental Setting

Baselines and Metrics. As very few dual-lens video inpainting methods are available, we compare our approach with representative methods for solving related tasks, including ten SOTA single-lens video inpainting methods, *i.e.*, FGVC (Gao et al. 2020), CPVINet (Lee et al. 2019), OPN (Seoung, Sungho et al. 2019), STTN (Zeng et al. 2020), FuseFormer (Liu et al. 2021), E2FGVI (Li et al. 2022b), ProPainter (Zhou et al. 2023), WaveFormer (Wu et al. 2024b), DiffuEraser (Li et al. 2025d), and VideoPainter (Bian et al. 2025) and the only dual-lens video inpainting method SVINet (Wu et al. 2023a). To ensure the comparability of results, we fine-tuned these baselines using their released models and codes on the same training data as our method, and we report their best results. Furthermore, we adopt PSNR (Haotian, Long et al. 2019), SSIM (Seoung, Sungho et al. 2019), LPIPS (Zhang, Isola et al. 2018), E_{warp} (Lai, Huang et al. 2018) and EPE (Hirschmuller 2008) to report quantitative results.

4.2 Results under standard systems

Datasets. To evaluate the effectiveness of our method under standard binocular systems, we follow the setting of SVINet and use the SVI (Wu et al. 2023a) as our benchmark dataset. The SVI dataset is currently the only inpainting dataset specifically designed for stereo videos captured by a standard binocular systems. It contains 785 stereo video pairs recorded by a standard binocular systems with same field of view (FoV) and optical zoom factor. The dataset is divided into three parts containing 350, 135 and 200 video pairs for training, validation and testing, respectively.

Quantitative Comparisons with SOTAs. The left of Tab. 1 shows the quantitative results of our method and other baselines on SVI dataset. As shown in table, the inpainting performance of our method is significantly better than existing single-lens baselines. These results further illustrate the necessity of developing dual-lens video inpainting algorithms. Furthermore, compared to SVINet, our method leverages the superior long-range modeling capabilities of the transformer structure to propagate available contents from distant frames into the missing regions. Therefore, our method obtains better inpainting performance than SVINet.

Qualitative Comparisons and Analyses. Fig. 3 presents three examples comparing our model with two competitive

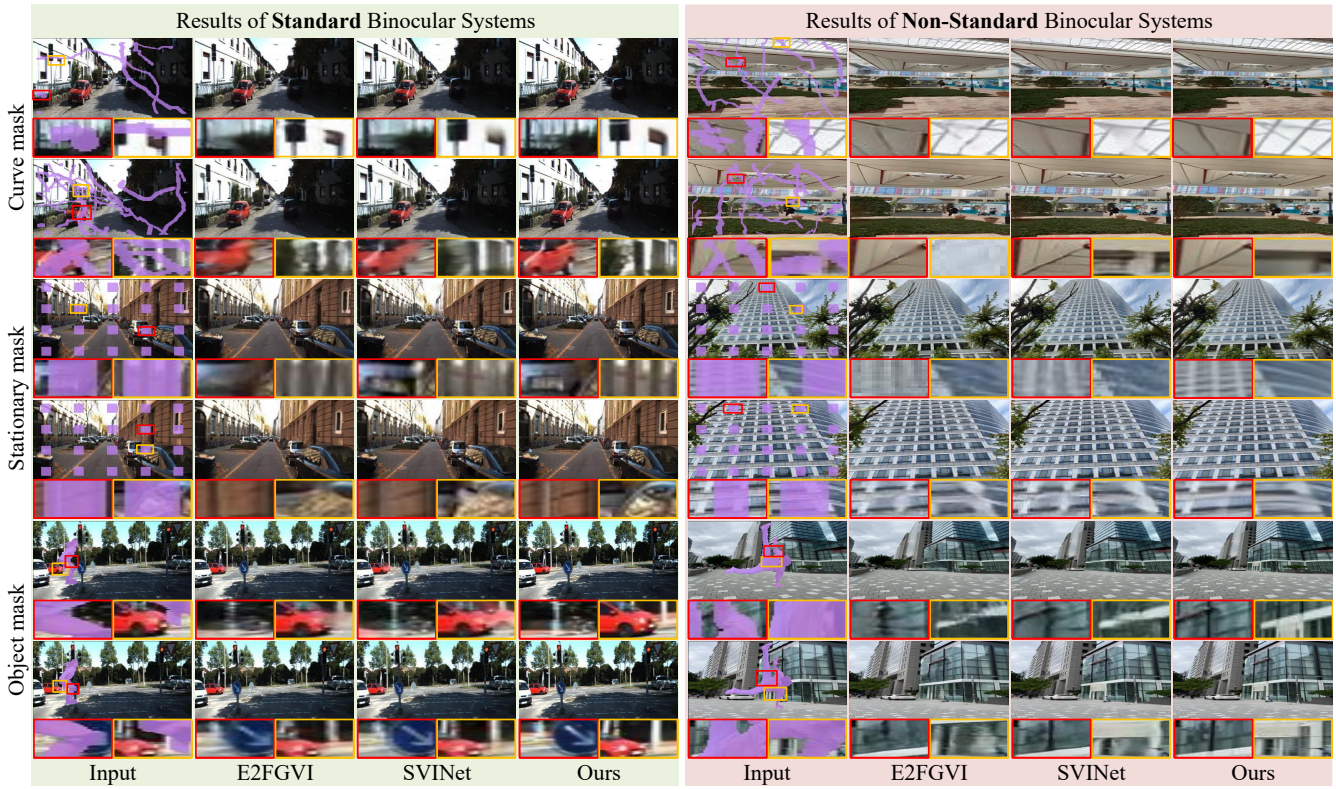

Figure 3: Qualitative results compared with SOTA baselines under three different mask settings. For standard (non-standard) binocular systems, the top row is the left (wide-angle) view, the bottom row is the right (telephoto) view.

baselines under curve, stationary, and object mask settings. As can be observed, the SOTA single-lens video inpainting model E2FGVI can generate specious missing contents for each lens, it fails to maintain the consistency between the lenses. Although the results by SVINet achieve better inter-lens consistency, but lacks necessary detail. In contrast, our model can not only generate vivid textures but also produce inter-lens consistent contents.

4.3 Results under non-standard systems

Datasets. To date, there is no publicly inpainting dataset specifically designed for video captured by a non-standard binocular systems. As a result, we customize a video inpainting dataset under non-standard binocular systems by using RealMCVSR (Lee et al. 2022) dataset, named DCVI. Similar to the synthetic method of SVI dataset, we use the original video in RealMCVSR dataset, recorded with telephoto and wide-angle lenses with different FoV and optical zoom factor, as the ground truth. Additionally, we utilize the method described in previous works (Ji et al. 2024; Wu et al. 2025a) to generate masks of the missing regions. The customized DCVI dataset includes 400 training videos, 42 verification videos and 94 test videos. Each video pair consists of a wide-angle view that provides a broader image display and a telephoto view that contains more local details. To the best of our knowledge, DCVI is the first dual-lens video inpainting dataset under non-standard binocular systems.

Quantitative Comparisons with SOTAs. The right of Tab. 1 presents the quantitative results of our method and other baselines on DCVI dataset. As shown in table, our method outperforms all baselines in terms of the all five metrics. The results demonstrate that our method can generate videos with less distortion (PSNR and SSIM), more visually plausible contents (LPIPS), superior temporal coherence (E_{warp}), and better inter-lens consistency (EPE).

Qualitative Comparisons and Analyses. To visually inspect the visual results, we provide three samples visual comparisons in the right of Fig 3. The results show that E2FGVI tends to generate inconsistent missing contents between lenses. Compared with E2FGVI, SVINet can explore the consistency cues between lenses, but fail to capture the visible contents of long-distance frames, resulting in blurry or unrealistic artifacts. In contrast, the proposed method can produce missing contents that is more spatially reasonable, temporally coherent, and more consistent between lenses.

4.4 Efficiency Analysis

We compare the efficiency of our method with E2FGVI and SVINet. As shown in Tab. 2, our method ranks first in inference speed a single Titan RTX GPU, significantly outperforming E2FGVI and SVINet. As for the FLOPs, our method is 676.15G lower than the E2FGVI and SVINet, which indicates that our method is highly efficient. Note that the values of E2FGVI in the table have been multiplied by 2

Methods	E2FGVI	SVINet	Ours
FLOPs(:G)	884.36	861.26	**676.15**
Time(:S)	0.52	0.42	**0.33**

Table 2: Efficiency analysis.

Methods	WaveFormer		Ours	
	Standard	Sparse	Standard	Sparse
PSNR↑	28.2999	28.4816	32.0241	**32.2054**
SSIM↑	0.9127	0.9273	0.9476	**0.9618**

Table 3: Effectiveness of SSTA.

for fair comparison.

4.5 Ablation Study

Effectiveness of SSTA. We conduct an ablation study on our SSTA. As shown in Fig. 4, we can observe that the model with *standard spatial-temporal attention* introduces excessive redundant contents into the missing region, resulting in blurred results, In contrast, the model with *sparse spatial-temporal attention (SSTA)* can generate more textures. Furthermore, Tab. 3 further verifies the effectiveness of SSTA on WaveFormer and our method. As shown in Tab. 3, the model with SSTA achieves better performance.

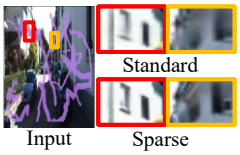

Figure 4: Ablation of SSTA.

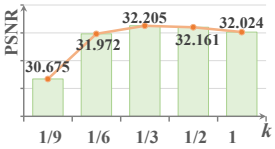

Figure 5: Effect of k.

Effect of the number of k**.** In the top-k adaptive selection strategy, the parameter k is employed to dynamically control the magnitude of sparsity. We investigates the influence of the parameter k on the inpainting performance. As shown in Fig. 5, the model achieves the best performance when only aggregates the features with correlations within top-$\frac{1}{3}$ range. Afterwards, with the continuous increase of k, the inpainting performance gradually decreases due to the introduction of irrelevant and useless features.

Effectiveness of MSFF. We also perform an ablation study on our MSFF module. As shown in Tab. 4, the single-scale feed-forward network using only 3×3 convolutional layers obtain the worst performance (PSNR=31.7341). With the increase of multi-scale information in the feed-forward network, the performance has been further improved. The model achieves the best performance (PSNR=32.2054), when simultaneously using 3×3, 5×5 and 7×7 convolutions in the feed-forward network. Such results suggest that rich multi-scale information can help improve the performance.

Effectiveness of CLTT. To verify the effectiveness of CLTT, we compare PAM, modified PAM (mPAM), and CLTT in SVINet and our method. As shown in Tab. 5, the model with CLTT achieves the best EPE across both methods. Furthermore, we compare the parallax flow of the model with modified PAM and CLTT in Fig. 6. From Fig. 6, we can be observed that the parallax flow between lenses using the CLTT module is closer to the ground-truth. This demonstrates that the CLTT can better model inter-lens consistency.

Conv. Size 3×3	5×5	7×7	PSNR↑	SSIM↑	E_{warp} ↓	LPIPS↓	EPE↓
✓			31.7341	0.9508	0.7357	0.3164	0.3296
✓	✓		32.0209	0.9553	0.7264	0.3118	0.3211
✓	✓	✓	**32.2054**	**0.9618**	**0.7039**	**0.2929**	**0.3109**

Table 4: Ablation study of MSFF.

Methods	SVINet			Ours		
	PAM	mPAM	CLTT	PAM	mPAM	CLTT
EPE↓	0.3725	0.3657	0.3372	0.3515	0.3497	**0.3109**

Table 5: Effectiveness of CLTT.

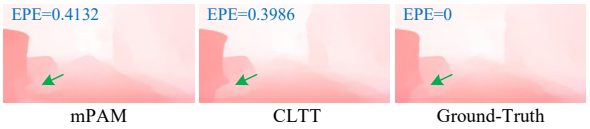

Figure 6: Parallax flow between CLTT and modified PAM.

Generalizability of CLTT. In Tab. 6, we integrate CLTT into ProPainter and WaveFormer to verify its generalization ability. As shown in Tab. 6, CLTT effectively enhances the inter-lens consistency of single-lens inpainting method in the dual-lens video inpainting task. This demonstrates the strong generalization ability and practical potential of CLTT.

Methods	PSNR↑	SSIM↑	E_{warp} ↓	LPIPS↓	EPE↓
ProPainter	28.0898	0.9043	0.8426	0.4192	0.5147
ProPainter+CLTT	29.2017	0.9183	0.7907	0.3524	0.3739
WaveFormer	28.2999	0.9127	0.8372	0.3776	0.5343
WaveFormer+CLTT	29.3950	0.9209	0.7895	0.3381	0.3606

Table 6: Generalizability of CLTT.

5 Conclusion

This paper proposes a novel framework named DLVINet for dual-lens video inpainting task in both standard and non-standard binocular systems. DLVINet consists of two core components, *i.e.*, the sparse spatial-temporal transformer and the cross-lens texture transformer. The former aims to generate missing contents for each lens by introducing the sparse spatial-temporal attention module and the multi-scale feed-forward module. The latter is designed to model the inter-lens consistency by capturing global correspondence across different lenses. Extensive experiments demonstrate the superior performance of our method on both videos captured by standard and non-standard binocular systems.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (U2336212, 62502447, 62402432), the Fundamental Research Funds for the Central Universities (226-2024-00058), the Natural Science Foundation of Zhejiang Province (LDT23F02023F02), the "Leading Goose" R&D Program of Zhejiang Province under Grant (2024C01101), the Fundamental Research Funds for the Zhejiang Provincial Universities (226-2024-00208), the Postdoctoral Fellowship Program of CPSF (GZC20251086), and the China Postdoctoral Science Foundation (2024M762830).

References

Bian, Y.; Zhang, Z.; Ju, X.; et al. 2025. VideoPainter: Any-length Video Inpainting and Editing with Plug-and-Play Context Control. *arXiv preprint arXiv:2503.05639*.

Cai, J.; Li, C.; Tao, X.; et al. 2022. DeViT: Deformed Vision Transformers in Video Inpainting. In *ACMMM*, 779–789.

Chang, Y.-L.; Liu, Z. Y.; Lee, K.-Y.; et al. 2019. Free-form video inpainting with 3D gated convolution and temporal patchGAN. In *ICCV*, 9066–9075.

Chen, C.; Qing, C.; Xu, X.; et al. 2022. Cross Parallax Attention Network for Stereo Image Super-Resolution. *IEEE TMM*, 24: 202–216.

Chen, K.; Wu, Z.; Hou, W.; et al. 2025. Prompt-Aware Controllable Shadow Removal. *arXiv preprint arXiv:2501.15043*.

Chen, S.; Ma, W.; Qin, Y.; et al. 2019. CNN-based stereoscopic image inpainting. In *ICIG*, 95–106.

Chen, T.; Cheng, Y.; Gan, Z.; et al. 2021. Chasing sparsity in vision transformers: An end-to-end exploration. In *NeurIPS*, volume 34, 19974–19988.

Cheng, M.; Ma, H.; Ma, Q.; et al. 2023. Hybrid Transformer and CNN Attention Network for Stereo Image Super-resolution. In *CVPR Workshops*, 1702–1711.

Cho, S.; Oh, S. W.; Lee, S.; et al. 2025. Elevating Flow-Guided Video Inpainting with Reference Generation. In *AAAI*, 2527–2535.

Fan, H.; Yang, Y.; Kankanhalli, M.; et al. 2021. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *CVPR*, 14204–14213.

Fan, H.; Yang, Y.; Kankanhalli, M.; et al. 2022. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE TPAMI*, 45(2): 2181–2192.

Gao, C.; Saraf, A.; Huang, J.-B.; et al. 2020. Flow-edge Guided Video Completion. In *ECCV*, 713–729.

Gu, J.; Li, K.; Wang, F.; et al. 2025. Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition. *arXiv preprint arXiv:2507.21977*.

Haotian, Z.; Long, M.; et al. 2019. An Internal Learning Approach to Video Inpainting. In *ICCV*, 2720–2729.

Hirschmuller, H. 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE TPAMI*.

Hou, J.; Ji, Z.; Yang, J.; et al. 2024. MCD-Net: toward RGB-D video inpainting in real-world scenes. *IEEE TIP*.

Hua, D.; Chen, Q.; Wu, Z.; et al. 2025. Perceptual Transform Fusion of Infrared and Visible Images. *IEEE TCSVT*.

Ji, Z.; Su, Y.; Zhang, Y.; et al. 2024. RAFormer: redundancy-aware transformer for video wire inpainting. *arXiv preprint arXiv:2404.15802*.

Kim, D.; Woo, S.; Lee, J.-Y.; et al. 2020. Recurrent Temporal Aggregation Framework for Deep Video Inpainting. *IEEE TPAMI*, 42(5): 1038–1052.

Lai, W.-S.; Huang, J.-B.; et al. 2018. Learning blind video temporal consistency. In *ECCV*, 179–195.

Lee, J.; Lee, M.; Cho, S.; et al. 2022. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 17824–17833.

Lee, M.; Cho, S.; Shin, C.; et al. 2025. Video diffusion models are strong video inpainter. In *AAAI*, 4526–4533.

Lee, S.; Oh, S. W.; Won, D.; et al. 2019. Copy-and-paste networks for deep video inpainting. In *ICCV*, 4413–4421.

Li, A.; Zhao, S.; Zhang, Q.; et al. 2022a. Iterative Geometry-Aware Cross Guidance Network for Stereo Image Inpainting. In *IJCAI*, 1053–1059.

Li, H.; Chen, X.; Li, M.; et al. 2023. Learning A Sparse Transformer Network for Effective Image Deraining. In *CVPR*, 5896–5905.

Li, K.; Guo, D.; Chen, G.; et al. 2025a. Prototypical calibrating ambiguous samples for micro-action recognition. In *AAAI*, 4815–4823.

Li, K.; Liu, P.; Guo, D.; et al. 2024a. Mmad: Multi-label micro-action detection in videos. *arXiv preprint arXiv:2407.05311*.

Li, Q.; Chen, Q.; Chen, H.; et al. 2025b. Progressive Large-Scale Modeling via Temporal-Spatial Focus Connector for Micro-Action Recognition. In *ACMMM*, 14222–14228.

Li, Q.; He, F.; Chen, H.; et al. 2025c. Unleashing Foundation Vision Models: Adaptive Transfer for Diverse Data-Limited Scientific Domains. In *NeurIPS*.

Li, Q.; Wang, Y.; Zhang, Y.; et al. 2024b. Fuzzy-vit: A deep neuro-fuzzy system for cross-domain transfer learning from large-scale general data to medical image. *IEEE TFS*.

Li, X.; Xue, H.; Ren, P.; et al. 2025d. DiffuEraser: A Diffusion Model for Video Inpainting. *arXiv preprint arXiv:2501.10018*.

Li, Z.; Lu, C.-Z.; Qin, J.; et al. 2022b. Towards an End-to-End Framework for Flow-Guided Video Inpainting. In *CVPR*, 17562–17571.

Liang, C.; Wang, W.; Miao, J.; et al. 2022. Gmmseg: Gaussian mixture based generative semantic segmentation models. *NeurIPS*, 35: 31360–31375.

Liang, C.; Wang, W.; Miao, J.; et al. 2023a. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *ICCV*, 16197–16208.

Liang, C.; Wang, W.; Zhou, T.; et al. 2023b. Local-global context aware transformer for language-guided video segmentation. *IEEE TPAMI*, 45(8): 10055–10069.

Liu, A.; Li, S.; Chang, Y.; et al. 2024a. Coarse-to-fine cross-view interaction based accurate stereo image super-resolution network. *IEEE TMM*, 26: 7321–7334.

Liu, H.; Wang, Y.; Qian, B.; et al. 2024b. Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting. In *CVPR*, 8038–8047.

Liu, R.; Deng, H.; Huang, Y.; et al. 2021. FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting. In *ICCV*, 14040–14049.

Liu, Y.; Xu, Y.; Wei, Y.; et al. 2025. Clear Nights Ahead: Towards Multi-Weather Nighttime Image Restoration. *arXiv preprint arXiv:2505.16479*.

Ma, W.; Zheng, M.; Ma, W.; et al. 2020. Learning across views for stereo image completion. *IET CVI*.

Quan, R.; Wu, Y.; Yu, X.; et al. 2021a. Progressive transfer learning for face anti-spoofing. *IEEE TIP*, 30: 3946–3955.

Quan, R.; Yu, X.; Liang, Y.; et al. 2021b. Removing raindrops and rain streaks in one go. In *CVPR*, 9147–9156.

Quan, W.; Chen, J.; Liu, Y.; et al. 2024. Deep learning-based image and video inpainting: A survey. *IJCV*.

Seoung, O., Wug; Sungho, L.; et al. 2019. Onion-Peel Networks for Deep Video Completion. In *ICCV*, 4402–4411.

Sun, H.; Li, Y.; Yang, K.; et al. 2025. VIP: Video Inpainting Pipeline for Real World Human Removal. *arXiv preprint arXiv:2504.03041*.

Tuo, Z.; Yang, H.; Fu, J.; et al. 2023. Learning Data-Driven Vector-Quantized Degradation Model for Animation Video Super-Resolution. In *ICCV*, 13133–13143.

Wang, C.; Huang, H.; Han, X.; et al. 2019. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, 5232–5239.

Wang, F.; Li, K.; Nie, Y.; et al. 2025. Exploiting ensemble learning for cross-view isolated sign language recognition. In *ACM WWW*, 2453–2457.

Wang, J.; Wu, Z.; Xuan, H.; et al. 2024. Text-Video Completion Networks With Motion Compensation And Attention Aggregation. In *ICASSP*, 2990–2994.

Wang, J.; Xuan, H.; Wu, Z.; et al. 2023. Semantic-guided completion network for video inpainting in complex urban scene. In *ICASSP*, 224–236.

Wang, L.; Guo, Y.; Wang, Y.; et al. 2022. Parallax Attention for Unsupervised Stereo Correspondence Learning. *IEEE TPAMI*, 44(4): 2108–2125.

Wu, J.; Li, X.; Si, C.; et al. 2024a. Towards Language-Driven Video Inpainting via Multimodal Large Language Models. In *CVPR*, 12501–12511.

Wu, Z.; Chen, K.; Li, K.; et al. 2025a. BVINet: Unlocking Blind Video Inpainting with Zero Annotations. *arXiv preprint arXiv:2502.01181*.

Wu, Z.; Li, K.; Fan, H.; et al. 2025b. Drafting and Revision: Advancing High-Fidelity Video Inpainting. In *IJCAI*.

Wu, Z.; Sun, C.; Xuan, H.; et al. 2023a. Deep Stereo Video Inpainting. In *CVPR*, 5693–5702.

Wu, Z.; Sun, C.; Xuan, H.; et al. 2023b. Divide-and-Conquer Completion Network for Video Inpainting. *IEEE TCSVT*.

Wu, Z.; Sun, C.; Xuan, H.; et al. 2024b. WaveFormer: Wavelet Transformer for Noise-Robust Video Inpainting. In *AAAI*, 6180–6188.

Wu, Z.; Xuan, H.; Sun, C.; et al. 2023c. Semi-Supervised Video Inpainting With Cycle Consistency Constraints. In *CVPR*, 22586–22595.

Xu, R.; Li, X.; Zhou, B.; et al. 2019. Deep flow-guided video inpainting. In *CVPR*, 3723–3732.

Xu, Y.; Sun, Y.; Yang, Z.; et al. 2022a. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *CVPR*, 14329–14339.

Xu, Y.; Yu, X.; Zhang, J.; et al. 2022b. Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting. *IEEE TIP*.

Xu, Y.; Zhu, L.; Yi, Y.; et al. 2025. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *ICCV*, 17675–17687.

Yang, S.; Gu, Z.; Hou, L.; et al. 2025. MTV-Inpaint: Multi-Task Long Video Inpainting. *arXiv preprint arXiv:2503.11412*.

Zeng, Y.; Fu, J.; Chao, H.; et al. 2020. Learning Joint Spatial-Temporal Transformations for Video Inpainting. In *ECCV*, 3723–3732.

Zhang, J.; Zhang, Y.; Gu, J.; et al. 2023a. Accurate Image Restoration with Attention Retractable Transformer. In *ICLR*.

Zhang, K.; Peng, J.; Fu, J.; et al. 2024a. Exploiting Optical Flow Guidance for Transformer-Based Video Inpainting. *IEEE TPAMI*, 1–16.

Zhang, R.; Isola, P.; et al. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, S.; Yu, W.; Jiang, F.; et al. 2024b. Stereo image restoration via attention-guided correspondence learning. *IEEE TPAMI*, 46(7): 4850–4865.

Zhang, Y.; Wu, Z.; Yan, Y.; et al. 2023b. Pfta-net: Progressive feature alignment and temporal attention fusion networks for video inpainting. In *ICIP*, 191–195.

Zhang, Z.; Wu, B.; Wang, X.; et al. 2024c. AVID: Any-Length Video Inpainting with Diffusion Model. In *CVPR*.

Zhao, G.; Lin, J.; Zhang, Z.; et al. 2020. Explicit sparse transformer: Concentrated attention through explicit selection. In *ICLR*.

Zhou, S.; Chen, D.; Pan, J.; et al. 2024. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *CVPR*, 2952–2963.

Zhou, S.; Li, C.; Chan, K. C.; et al. 2023. ProPainter: Improving Propagation and Transformer for Video Inpainting. In *ICCV*, 10477–10486.

Zi, B.; Zhao, S.; Qi, X.; et al. 2025. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In *AAAI*.

Zou, X.; Yang, L.; Liu, D.; et al. 2021. Progressive Temporal Feature Alignment Network for Video Inpainting. In *CVPR*, 16448–16457.