Drafting and Revision: Advancing High-Fidelity Video Inpainting

Zhiliang Wu, **Kun Li**, **Hehe Fan** and **Yi Yang**[†]
ReLER, CCAI, Zhejiang University, China

Abstract

Video inpainting aims to fill the missing regions in video with spatial-temporally coherent contents. Existing methods usually treat the missing contents as a whole and adopt a hybrid objective containing a reconstruction loss and an adversarial loss to train the model. However, these two kinds of loss focus on contents at different frequencies, simply combining them may cause inter-frequency conflicts, leading the trained model to generate compromised results. Inspired by the common corrupted painting restoration process of "drawing a draft first and then revising the details later", this paper proposes a Drafting-and-Revision Completion Network (DRCN) for video inpainting. Specifically, we first design a Drafting Network that utilizes the temporal information to complete the low-frequency semantic structure at low resolution. Then, a Revision Network is developed to hallucinate high-frequency details at high resolution by using the output of Drafting Network. In this way, adversarial loss and reconstruction loss can be applied to high-frequency and low-frequency respectively, effectively mitigating inter-frequency conflicts. Furthermore, Revision Network can be stacked in a pyramid manner to generate higher resolution details, which provide a feasible solution for high-resolution video inpainting. Experiments show that DRCN achieves improvements of 7.43% and 12.64% in E_{warp} and LPIPS, and can handle higher resolution videos on limited GPU memory.

1 Introduction

Video inpainting aims to fill the missing regions of a video with spatial-temporally coherent contents, which is a fundamental visual restoration task. High-quality video inpainting can benefit general users in various applications, such as object removal [Wu *et al.*, 2023c], video restoration [Wang *et al.*, 2024], autonomous driving [Zhang *et al.*, 2023], and so on. Unlike image inpainting [Liu *et al.*, 2024a; Zhuang *et al.*, 2024], which primarily focuses on the spatial dimension, video inpainting pays more attention to the temporal information. Directly using image inpainting methods to individual frames for video inpainting will neglect the motion continuity between frames, resulting in flicker artifacts.

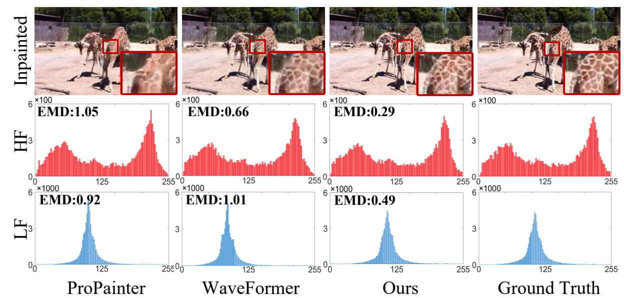

Figure 1: Results comparison of ProPainter [Zhou *et al.*, 2023], WaveFormer [Wu *et al.*, 2024], and our method. Due to frequency conflict, ProPainter and WaveFormer fail to generate the missing details. In contrast, our method successfully generates richer and more realistic textural details. EMD denotes the Earth Mover's Distance [Rubner *et al.*, 2000] between the ground-truth histogram and the result histogram in the low-frequency (LF)/high-frequency (HF), where a lower value indicates better result.

Recently, several deep learning-based video inpainting methods [Li *et al.*, 2020; Liu *et al.*, 2021; Zhang *et al.*, 2022c; Wang *et al.*, 2023; Wu *et al.*, 2021] have been proposed and have achieved significant results. Nevertheless, these methods always treat the missing regions as a whole and employ a hybrid objective consisting of a reconstruction loss (L1/L2 norm) and an adversarial loss to train the model, resulting in over-smooth generated missing contents compared to reveal realistic detail, as illustrated in Fig. 1. On the one hand, these methods treat the missing regions as a whole, *i.e.*, all pixels are viewed equally. They do not distinguish between flat regions and texture details, which are contained in low-frequency and high-frequency components respectively. In this way, the trained models will be easily dominated by flat regions which are the most common [Wu *et al.*, 2023b]. On the other hand, the reconstruction loss and the adversarial loss tend to synthesis contents at different frequencies, *i.e.*, the former focuses on recovering the low-frequency global structures [Pathak *et al.*, 2016], while the latter prefers to generate the high-frequency texture details [Yu *et al.*, 2021]. Simply combining these two losses may cause inter-frequency conflicts, leading to much less favourable results.

[†]Corresponding author.

Recall the process of a painter inpainting a corrupted painting, we can find that a common practice, especially for a beginner, is to draw a draft first to capture the global structure of the painting, and then gradually revise the local details based on the global structure, rather than directly completing the final inpainting part-by-part. Inspired by such a "*drawing a draft first and revising the details later*" manner [Lin *et al.*, 2021], we propose a novel **D**rafting-and-**R**evision **C**ompletion **N**etwork (**DRCN**) for video inpainting. DRCN decomposes the video frame into low-frequency and high-frequency components, and designs the a *Drafting Network* and a *Revision Network* to complete them respectively. In this way, we can not only avoid inter-frequency conflicts by applying adversarial loss and L1 loss to the high-frequency and low-frequency branches separately, but also solve the problem of varying difficulties in generating low-frequency semantics and high-frequency details.

Specifically, we first adopt Laplacian transform to decompose the frames into low-frequency and high-frequency components. By doing this, flat regions common in frame are recorded in low-resolution low-frequency components, while texture details are mainly concentrated in high-resolution high-frequency components. Next, we develop a *Drafting Network* to complete the semantic structure of missing regions in low-frequency components at low resolution, benefiting from its larger receptive field and less local details. Thereafter, a *Revision Network* is designed to revise the high-frequency local details of the missing contents at $2\times$ resolution. The Revision Network utilizes the draft generated by Drafting Network to guide the high-frequency components in generating the missing contents. Finally, the completed low-frequency and high-frequency components are aggregated to yield the final inpainting result by inverse Laplacian transform. Notably, our Revision Network can be stacked in a pyramid manner to complete high-frequency details at higher resolution, which can provide a feasible solution for the high-resolution video inpainting. Extensive experimental results demonstrate that DRCN can generate the missing contents with richer textures compared to baselines.

To sum up, our contributions are summarized as follows:

- A novel Drafting-and-Revision Completion Network (DRCN) is designed to effectively mitigate the inter-frequency conflict in video inpainting.
- A feasible high-resolution video inpainting solution is first attempted. Our network can handle higher resolution videos on limited GPU memory.
- Extensive experiments on two benchmark datasets, including Youtube-vos [Xu *et al.*, 2018] and DAVIS [Perazzi *et al.*, 2016], demonstrate the superiority of our DRCN in both quantitative and qualitative evaluations.

2 Related Work

Video Inpainting. Recently, several deep learning based video inpainting methods have been proposed and achieved great progress. According to the network architectures involved, these methods can be summarised into three groups.

3D CNNs-based Methods: Some researchers [Chang *et al.*, 2019; Kim *et al.*, 2019] utilize 3D CNNs to integrate spatial-temporal information and fill in missing regions. Although they have produced promising results, the computational complexity of 3D CNNs is relatively higher, which limits their practical application [Wu *et al.*, 2023a].

Optical Flow-based Methods: Unlike 3D CNNs-based methods, optical flow-based methods [Xu *et al.*, 2019; Gao *et al.*, 2020; Zhang *et al.*, 2024] formulated the video inpainting as a flow-guided pixel propagation task. They first completed the optical flow by a flow completion network, and then propagated the relevant pixels using the completed flow into missing regions. Despite achieving encouraging results, they still suffer from challenges in propagating valid pixels from distant frames. In a sense, their performance significantly decrease when the missing regions are large and slow-moving [Zhou *et al.*, 2023].

Attention-based Methods: Attention [Li *et al.*, 2023; Li *et al.*, 2025a; Wang *et al.*, 2025; Liu *et al.*, 2024b; Li *et al.*, 2024; Li *et al.*, 2025b] has been proven to model long-distance dependencies, some methods [Liu *et al.*, 2021; Zhang *et al.*, 2022b; Zhou *et al.*, 2023; Wu *et al.*, 2024] incorporated attention mechanism to extend the limited temporal receptive field. These methods retrieve relevant information from long-distance frames by this mechanism and adopted weighted operation to generate missing contents. Among these methods, Zeng et al. [Zeng *et al.*, 2020], Liu et al. [Liu *et al.*, 2021], Zhang et al. [Zhang *et al.*, 2024], and Wu et al. [Wu *et al.*, 2024] employed transformers to retrieve similar features in a considerable temporal receptive field, resulting in high-quality video inpainting.

In spite of the promising results shown by these methods, over-smoothed missing contents are generated, failing to infer realistic details. Meanwhile, these methods can only handle low resolution videos (typically smaller than 1K) due to constraints in GPU memory and computation time, and are ineffective for high-resolution videos in real-world scenarios.

Coarse-to-Fine Strategy. In the field of visual restoration, the coarse-to-fine strategy has been proposed and utilized in various tasks, such as image super-resolution [Tian *et al.*, 2021], video super-resolution [Xiao *et al.*, 2023], and so on. Although these methods have achieved remarkable results by coarse-to-fine strategy, they usually treat the content as a whole and adopt a hybrid objective consisting of reconstruction loss and adversarial loss to train the network, leading the trained model to generate compromised results.

Unlike the coarse-to-fine approach, our drafting-and-revision strategy employs the Laplace transform to decompose the frame into low-resolution low-frequency components and high-resolution high-frequency ones, and processes them separately in the drafting and revision stages. Such a design not only avoids inter-frequency conflicts by applying adversarial loss and L1 loss to different frequency branches separately, but also enables stacking the Revision Networks in a pyramid manner to process high-resolution video.

3 Proposed Method

3.1 Formulation and Overview

Problem Formulation. Assume $\boldsymbol{X} = \{\boldsymbol{x}_i \in \mathbb{R}^{h\times w\times 3}\}_1^T$ is a corrupted video with length T. The binary mask $\boldsymbol{M} =$

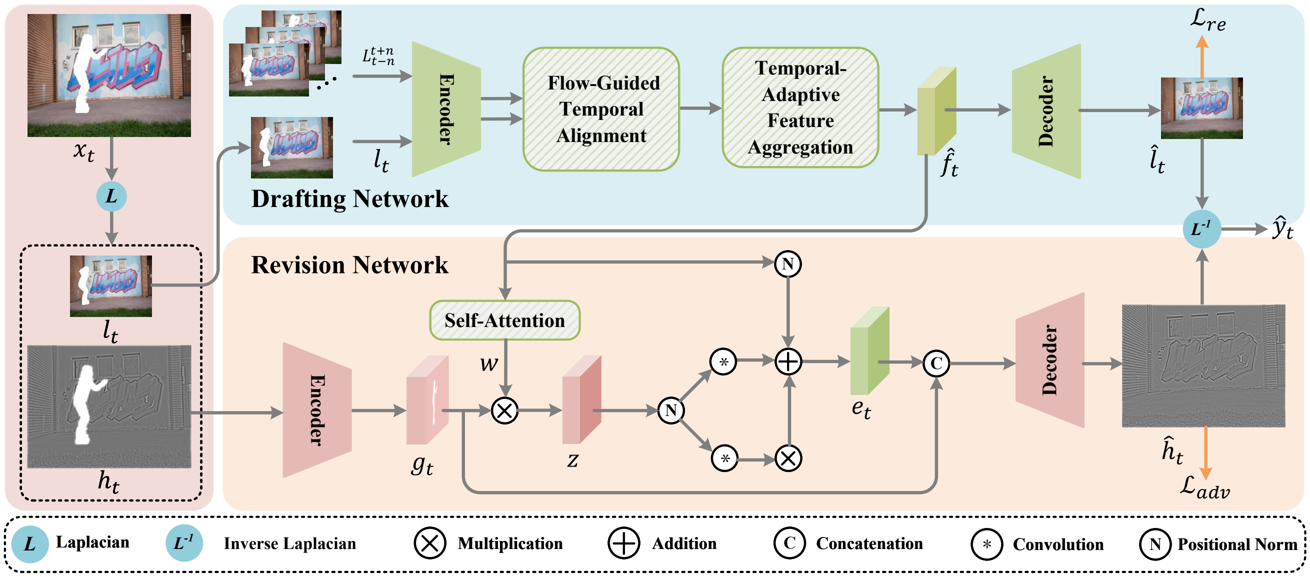

Figure 2: Overview of our framework. We first generate image pyramid $\{\boldsymbol{l}_i, \boldsymbol{h}_i\}_{t-n}^{t+n}$ from video frames $\{\boldsymbol{x}_i\}_{t-n}^{t+n}$ by using Laplacian transform. Then, ***Drafting Network*** generates rough low-resolution completed result $\widehat{\boldsymbol{l}}_t$, which has complete semantics but lacks detailed information. Next, ***Revision Network*** completes high-frequency residual $\boldsymbol{h}_t$ at high-resolution to obtain the completed high-frequency component $\widehat{\boldsymbol{h}}_t$. Finally, the final inpainting result $\widehat{\boldsymbol{y}}_t$ is obtained by aggregating the pyramid outputs $\widehat{\boldsymbol{l}}_t$ and $\widehat{\boldsymbol{h}}_t$. Remarkably, in our framework, the adversarial (or L1) loss is applied to the high-frequency (or low-frequency) branch separately to mitigate the inter-frequency conflicts.

$\{\boldsymbol{m}_i \in \mathbb{R}^{h\times w\times 1}\}_1^T$ denotes the missing regions of corresponding frames. For each mask $\boldsymbol{m}_i$, "0" indicates that the valid region of $\boldsymbol{x}_i$, and "1" stands for the missing regions. The goal of video inpainting is to generate a completed video $\widehat{\boldsymbol{Y}} = \{\widehat{\boldsymbol{y}}_i \in \mathbb{R}^{h\times w\times 3}\}_1^T$, which should be consistent with ground truth video $\boldsymbol{Y} = \{\boldsymbol{y}_i \in \mathbb{R}^{h\times w\times 3}\}_1^T$ in both spatial and temporal dimensions.

In practice, we usually use a deep neural network $\mathcal{DNN}(\cdot)$ to predict $\widehat{\boldsymbol{y}}_t$ frame by frame *i.e.*, $\widehat{\boldsymbol{y}}_t = \mathcal{DNN}(\boldsymbol{x}_t, \boldsymbol{X}_{t-n}^{t+n}, \boldsymbol{m}_t)$. Here, $\boldsymbol{x}_t$ is the current frame that needs to be inpainted, called *target frame*. $\boldsymbol{X}_{t-n}^{t+n} = \{\boldsymbol{x}_{t-n}, \dots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}, \dots, \boldsymbol{x}_{t+n}\}$ denotes a short clip of neighboring frames with a center moment t and a temporal radius n, namely *reference frames*. Existing methods [Yu *et al.*, 2019; Zeng *et al.*, 2020; Liu *et al.*, 2021; Zhang *et al.*, 2022b; Zhou *et al.*, 2023; Wu *et al.*, 2024] always treat the missing contents as a whole and employ a hybrid objective $\mathcal{L}_{hy}$ to train the $\mathcal{DNN}(\cdot)$,

$$\mathcal{L}_{hy} = \lambda_{re}\mathcal{L}_{re} + \mathcal{L}_{adv}, \tag{1}$$

where $\mathcal{L}_{re}$ is the reconstruction (L1 or L2) loss, $\mathcal{L}_{adv}$ is the adversarial loss, and λ_{re} is the trade-off parameter.

We argue that these methods treat all pixels equally, resulting in the completed regions being too smooth and lacking realistic details. On the one hand, the reconstruction objectives of the missing regions are not consistent regarding different low-level frame elements, *e.g.*, smoothness preserving for flat regions, sharpening for edges and textures. On the other hand, the reconstruction loss $\mathcal{L}_{re}$ prefers to focus on low-frequency global structures [Deng *et al.*, 2019; Yu *et al.*, 2021], while the adversarial loss $\mathcal{L}_{adv}$ tends to concentrate on high-frequency texture details [Pathak *et al.*, 2016]. Simply combining them like Eq.(1) will lead to inter-frequency conflicts. To alleviate such conflicts, existing methods [Yu *et al.*, 2019; Zeng *et al.*, 2020; Liu *et al.*, 2021; Zhang *et al.*, 2022b; Zhou *et al.*, 2023; Wu *et al.*, 2024] attempt to balance the $\mathcal{L}_{re}$ and $\mathcal{L}_{adv}$ by adjusting the parameter λ_{re}. However, the missing contents in spatial domain are still generated with mixed frequency [Yu *et al.*, 2021]. Not only is this strategy sub-optimal, but adjusting hyper-parameters λ_{re} is trivial. Therefore, separate treatment of low-frequency and high-frequency of missing regions and explicitly applying the $\mathcal{L}_{adv}$ and $\mathcal{L}_{re}$ losses to different branches, is necessary to generate the missing contents with more realistic details.

Network Design. In this paper, we propose a Drafting-and-Revision Completion Network, named as DRCN. The pipeline of our DRCN is illustrated in Fig. 2. Given the target frame $\boldsymbol{x}_t$ and the reference frame $\boldsymbol{x}_r \in \boldsymbol{X}_{t-n}^{t+n}$, we first decompose them into low-frequency component $\boldsymbol{l}_t, \boldsymbol{l}_r \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$ and high-frequency component $\boldsymbol{h}_t, \boldsymbol{h}_r \in \mathbb{R}^{h\times w\times 3}$ by Laplacian transform. Notably, $\boldsymbol{l}_t$ and $\boldsymbol{l}_r$ mainly record the global semantic structure, while $\boldsymbol{h}_t$ and $\boldsymbol{h}_r$ generally contain their corresponding detailed information, *e.g.*, edges and textures. Then, the low-frequency component $\boldsymbol{l}_t$ and high-frequency component $\boldsymbol{h}_t$ are fed into ***Drafting Network*** and ***Revision Network*** to conduct the completion, respectively. Finally, the completed low-frequency and high-frequency components are aggregated to generate final inpainting result $\widehat{\boldsymbol{y}}_t \in \mathbb{R}^{h\times w\times 3}$ by the inverse Laplace transform. In this way, L1 loss $\mathcal{L}_{re}$ and adversarial loss $\mathcal{L}_{adv}$ can be applied to low-frequency and high-frequency components independently, effectively mitigating inter-frequency conflicts. Furthermore, Revision Network can be stacked in a pyramid manner to inpaint high-resolution video. In the following, we will introduce the Drafting Network and Revision Network in detail.

3.2 Drafting Network

At low resolution, the semantic structure is easier to complete due to the large receptive field and less local details [Wu *et al.*, 2023b]. Based on this fact, we design a Drafting Network to complete the semantic structures of the missing regions at low resolution. As shown in Fig. 2, our Drafting Network is built upon an encoder-decoder architecture, which consists of a frame-level encoder, a temporal alignment module, a feature aggregation module and a frame-level decoder. Temporal alignment and feature aggregation modules are the core components of the Drafting Network. The former performs the feature alignment to eliminate image variations between the reference frame and the target frame, while the latter aggregates the aligned features of the reference frame to generate the missing contents of the target frame.

Self-Supervised Flow-Guided Temporal Alignment. Since the movement of the camera or object cause the image variation, it is difficult to directly utilize the reference frames to complete the the missing regions of target frame. Therefore, an extra alignment module is necessary to eliminate the image variation between the reference frame and the target frame. Benefit from the capability of deformable convolution (DCN) [Dai *et al.*, 2017] to handle complex geometric transformations, some works [Wang *et al.*, 2019; Tian *et al.*, 2020; Wu *et al.*, 2023b] have proposed various forms of DCN-based temporal alignment module to achieve alignment between reference frame and target frame. Although these modules achieve excellent alignment results, they often suffer from offset overflow during training, thereby having a negative impact on model performance. To alleviate this problem, researchers [Chan *et al.*, 2022; Zhang *et al.*, 2022a; Liang *et al.*, 2022; Wu *et al.*, 2023a] used the optical flow between reference frame and target frame as base offset of DCN to train the network. However, these alignment module still face the following challenges:

- They typically relied on a heavyweight pre-trained DNN (*e.g.*, PWC-Net [Sun *et al.*, 2018]) to generate the optical flow, which significantly increases the computational cost, thus limiting their practical application.
- They were difficult to converge due to the lack of constraints during the training.

In fact, the optical flow in these alignment module only serves as a base offset to guide the training of DCN, which means that a coarse-grained optical flow is sufficient for the requirements [Zhang *et al.*, 2022a; Wu *et al.*, 2023a]. In other words, precise optical flow generated by the heavyweight DNN is redundant for these alignment module. Furthermore, target frames can act as labels to force the reference frames closer to them in a self-supervised manner during alignment, achieving fast convergence of the model.

For this purpose, we design a self-supervised flow-guided temporal alignment module. Specifically, we estimate the optical flow using a lightweight motion estimator that stacks 3 convolutional layers. Such a design not only significantly reduces the computational cost, but also allows the network to be trained from scratch to generate more suitable optical flows for video inpainting. Besides, we introduce an alignment loss $\mathcal{L}_a$ as a self-supervised constraint.

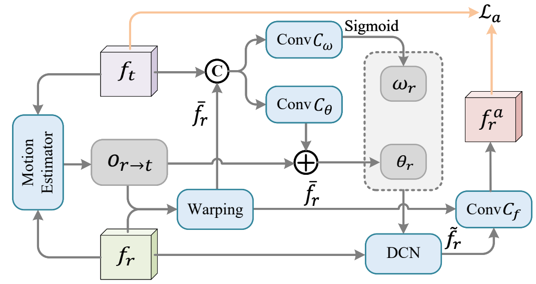

Figure 3: Illustration of flow-guided temporal alignment module.

As shown in Fig. 3, for the features $\boldsymbol{f}_r$ and features $\boldsymbol{f}_t$ corresponding to $\boldsymbol{l}_r$ and $\boldsymbol{l}_t$ are extracted by the frame-level encoder, we first estimate the optical flow $\boldsymbol{o}_{r\to t}$ by a lightweight motion estimator $\mathcal{M}(\cdot)$, and generate warped features $\bar{\boldsymbol{f}}_r$ by a warping operation [Wang *et al.*, 2019] $\mathcal{W}(\cdot)$,

$$\boldsymbol{o}_{r\to t} = \mathcal{M}(\boldsymbol{f}_r, \boldsymbol{f}_t), \tag{2}$$

$$\bar{\boldsymbol{f}}_r = \mathcal{W}(\boldsymbol{f}_r, \boldsymbol{o}_{r\to t}). \tag{3}$$

Then, the warped features $\bar{\boldsymbol{f}}_r$ and the features $\boldsymbol{f}_t$ are used to compute the offsets θ_r and modulation masks ω_r,

$$\theta_r = \boldsymbol{o}_{r\to t} + \mathcal{C}_\theta([\boldsymbol{f}_t, \bar{\boldsymbol{f}}_r]), \tag{4}$$

$$\omega_r = \sigma(\mathcal{C}_\omega([\boldsymbol{f}_t, \bar{\boldsymbol{f}}_r])), \tag{5}$$

where $\mathcal{C}_\theta(\cdot)$ and $\mathcal{C}_\omega(\cdot)$ presents 2D convolution, $[\cdot,\cdot]$ denotes the concatenation, and $\sigma(\cdot)$ indicates sigmoid function. Notably, when calculating the offsets θ_r based on Eq.(4), we consider the residuals of optical flow $\boldsymbol{o}_{r\to t}$ as a base offset of DCN instead of directly computing the offsets. Such a strategy can effectively mitigate the offset overflow [Chan *et al.*, 2022; Wu *et al.*, 2023d] of DCN during the training.

Next, the aligned reference frame features $\widetilde{\boldsymbol{f}}_r$ can be acquired by a DCN layer $\mathcal{DCN}(\cdot)$:

$$\widetilde{\boldsymbol{f}}_r = \mathcal{DCN}(\boldsymbol{f}_r; \theta_r, \omega_r). \tag{6}$$

Finally, to obtain a more robust alignment feature, we fuse $\bar{\boldsymbol{f}}_r$ calculated by Eq.(3) and $\widetilde{\boldsymbol{f}}_r$ calculated by Eq.(6) to generate the final aligned features $\boldsymbol{f}_r^a$,

$$\boldsymbol{f}_r^a = \mathcal{C}_f(\widetilde{\boldsymbol{f}}_r, \bar{\boldsymbol{f}}_r), \tag{7}$$

where $\mathcal{C}_f(\cdot)$ denotes feature-level fusion operation consisting of convolutional layers.

Temporal-Adaptive Feature Aggregation. Due to occlusion, blurry regions and parallax problems, different reference frames are not equally beneficial for reconstructing the missing contents [Wu *et al.*, 2023a; Chen *et al.*, 2025]. To solve this issue, we construct a temporal-adaptive feature aggregation module to generate the missing semantic structures.

Specifically, we first compute the attention weight $\boldsymbol{s}_r$ of the aligned reference frame features $\boldsymbol{f}_r^a$ by a softmax function:

$$\boldsymbol{s}_r = \frac{exp\left(\mathcal{C}_q(\boldsymbol{f}_t)^T \cdot \mathcal{C}_k(\boldsymbol{f}_r^a)\right)}{\sum_r exp\left(\mathcal{C}_q(\boldsymbol{f}_t)^T \cdot \mathcal{C}_k(\boldsymbol{f}_r^a)\right)}, \tag{8}$$

where $\mathcal{C}_q(\cdot)$ and $\mathcal{C}_k(\cdot)$ denote 1×1 2D convolution. Then, the attention-modulated features $\boldsymbol{q}_r$ can be obtained by $\boldsymbol{q}_r =$

$\mathcal{C}_v(\boldsymbol{f}_r^a) \odot \boldsymbol{s}_r$, where $\mathcal{C}_v(\cdot)$ indicates 1×1 2D convolution and $\odot$ presents the element-wise multiplication.

After obtaining the all attention-modulated features $\boldsymbol{q}_r$, $r \in \{t-n, \ldots, t-1, t+1, \ldots, t+n\}$, the completed features $\widehat{\boldsymbol{f}}_t$ can be generated by a aggregation convolutional layer $\mathcal{C}_a(\cdot)$,

$$\widehat{\boldsymbol{f}}_t = \mathcal{C}_a([\boldsymbol{q}_{t-n}, \ldots, \boldsymbol{q}_{t-1}, \boldsymbol{q}_{t+1}, \ldots, \boldsymbol{q}_{t+n}, \boldsymbol{f}_t, \boldsymbol{m}_t]). \quad (9)$$

Here, the size of the mask $\boldsymbol{m}_t$ is resized to fit the size of the features $\boldsymbol{f}_t$. The final completed low-frequency component $\widehat{\boldsymbol{l}}_t$ can be obtained by decoding $\widehat{\boldsymbol{f}}_t$ with the frame-level decoder.

3.3 Revision Network

The low-frequency component $\widehat{\boldsymbol{l}}_t$ generated by Drafting Network contains the complete semantic structure but lacks detailed information, *e.g.*, edges and textures. Adding the high-frequency information to the $\widehat{\boldsymbol{l}}_t$ will produce clearer result with richer details. Therefore, it is quite necessary to design a Revision Network to complete the high-frequency component $\boldsymbol{h}_t$. A naive design of Revision Network is to directly replicate our Drafting Network. Nevertheless, such a design will consume a lot of GPU memory since Revision Network needs to search for relevant information from multiple reference frames simultaneously, which is extremely detrimental to high-resolution video inpainting. For this purpose, we develop a Revision Network that exploits the inpainted low-frequency component $\widehat{\boldsymbol{l}}_t$ to guide the completion of the high-frequency component $\boldsymbol{h}_t$. In the real implementation, due to the high sparsity of the $\boldsymbol{h}_t$, directly summing or concatenating $\widehat{\boldsymbol{l}}_t$ and $\boldsymbol{h}_t$ to generate $\widehat{\boldsymbol{h}}_t$ will greatly suppress the high-frequency information. Therefore, aligning the $\widehat{\boldsymbol{l}}_t$ with the $\boldsymbol{h}_t$ is a crucial step in generating consistent and realistic high-frequency missing contents.

Specifically, we first encode the $\boldsymbol{h}_t$ into the feature $\boldsymbol{g}_t$ by the frame-level encoder, where $\boldsymbol{g}_t = \{\boldsymbol{g}_t^1, \ldots, \boldsymbol{g}_t^e\} \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times c}$ and $e = \frac{h}{2} \times \frac{w}{2}$. Then, the self-attention score $w_{i,j}$ of the feature $\widehat{\boldsymbol{f}}_t$ obtained by Eq.(9) is calculated as follows,

$$w_{i,j} = \frac{exp(\widehat{\mathcal{C}}_k(\widehat{\boldsymbol{f}}_t^i)^T \cdot \widehat{\mathcal{C}}_q(\widehat{\boldsymbol{f}}_t^j))}{\sum_i exp(\widehat{\mathcal{C}}_k(\widehat{\boldsymbol{f}}_t^i)^T \cdot \widehat{\mathcal{C}}_q(\widehat{\boldsymbol{f}}_t^j))}, \quad (10)$$

where $1 \leq i, j \leq e$, $\widehat{\mathcal{C}}_k(\cdot)$ and $\widehat{\mathcal{C}}_q(\cdot)$ denote 1×1 2D convolutions. The acquired attention map $\boldsymbol{w}$ depicts the correlation among the completed low-frequency feature. We aggregate the high-frequency feature of the valid regions to reconstruct the missing contents of $\boldsymbol{g}_t$ by

$$\boldsymbol{z}_i = \textstyle\sum_j w_{i,j} \cdot \widehat{\mathcal{C}}_v(\boldsymbol{g}_t^i), \quad (11)$$

where $\boldsymbol{z}_i$ is i-th aggregation feature and $\widehat{\mathcal{C}}_v(\cdot)$ is a 1×1 2D convolution layer.

Due to the sparseness of high-frequency feature, the magnitude of aggregation feature in Eq.(11) is relatively small. Inspired by frequency region attentive normalization [Yu *et al.*, 2021], we employ the parameter-free positional normalization [Li *et al.*, 2019] to normalize $\boldsymbol{z}$ while the preserving structural information. Similarly, the parameter-free positional normalization is also applied to $\widehat{\boldsymbol{f}}_t$. The aligned low-frequency feature $\boldsymbol{e}_t$ is computed as follows,

$$\boldsymbol{e}_t = \mathcal{C}_\gamma(\boldsymbol{z})\frac{\widehat{\boldsymbol{f}}_t - \mu_f}{\sigma_f} + \mathcal{C}_\beta(\boldsymbol{z}), \quad (12)$$

where μ_f and σ_f are the mean and standard deviation of $\widehat{\boldsymbol{f}}_t$ along the channel dimension, respectively. $\mathcal{C}_\gamma(\cdot)$ and $\mathcal{C}_\beta(\cdot)$ denote convolution operation. Finally, the completed high-frequency feature $\widehat{\boldsymbol{g}}_t$ is generated by a 1×1 2D convolution layer $\mathcal{C}_g(\cdot)$, *i.e.*, $\widehat{\boldsymbol{g}}_t = \mathcal{C}_g(\boldsymbol{e}_t, \boldsymbol{g}_t)$.

Data	Methods	PSNR↑	SSIM↑	E_{warp} ↓	LPIPS↓
Youtube-vos	VINet	26.174	0.8502	0.1694	1.0706
	FGVC	24.244	0.8114	0.2484	1.5884
	E2FGVI	30.064	0.9004	0.1490	0.5321
	FGT	30.811	0.9258	0.1308	0.4565
	STTN	28.993	0.8761	0.1523	0.6965
	FuseFormer	29.765	0.8876	0.1463	0.5481
	CPVINet	28.534	0.8798	0.1613	0.8126
	ProPainter	29.906	0.9050	0.1458	0.4962
	WaveFormer	33.264	0.9435	0.1184	0.2933
	Ours	**33.658**	**0.9532**	**0.1096**	**0.2565**
DAVIS	VINet	29.149	0.8965	0.1846	0.7262
	FGVC	28.936	0.8852	0.2122	0.9598
	E2FGVI	31.941	0.9188	0.4579	0.6344
	FGT	32.742	0.9272	0.1669	0.4240
	STTN	28.891	0.8719	0.1844	0.8683
	FuseFormer	29.627	0.8852	0.1767	0.6706
	CPVINet	30.234	0.8997	0.1892	0.6560
	ProPainter	31.967	0.9250	0.1655	0.4370
	WaveFormer	34.169	0.9475	0.1504	0.3137
	Ours	**34.676**	**0.9582**	**0.1287**	**0.2868**

Table 1: Quantitative results on Youtube-vos and DAVIS datasets.

4 Experiments

4.1 Experimental Setting

Datasets. Following previous works [Ren *et al.*, 2022; Zhou *et al.*, 2023; Wu *et al.*, 2024], two most commonly used datasets (Youtube-vos [Xu *et al.*, 2018] and DAVIS [Perazzi *et al.*, 2016]) are considered to verify the effectiveness of our method. The Youtube-vos [Xu *et al.*, 2018] dataset contains 4453 videos with various scenes, and is split into three parts containing 3471, 474 and 508 videos for training, validation and testing, respectively. As for the DAVIS [Perazzi *et al.*, 2016] dataset, it contains 150 high-quality videos of challenging motion-blur and appearance motions. Consistent with existing studies [Zhou *et al.*, 2023; Yu *et al.*, 2023; Zhang *et al.*, 2024; Wu *et al.*, 2024], 60 videos are used for training and 90 videos are utilized for testing.

Baselines and Evaluation Metrics. We select nine recently video inpainting methods as our baselines, including VINet [Kim *et al.*, 2019], CPVINet [Lee *et al.*, 2019], FGVC [Gao *et al.*, 2020], STTN [Zeng *et al.*, 2020], FuseFormer [Liu *et al.*, 2021], E2FGVI [Li *et al.*, 2022], FGT [Zhang *et al.*, 2022b], ProPainter [Zhou *et al.*, 2023], and WaveFormer [Wu *et al.*, 2024]. To ensure the fairness

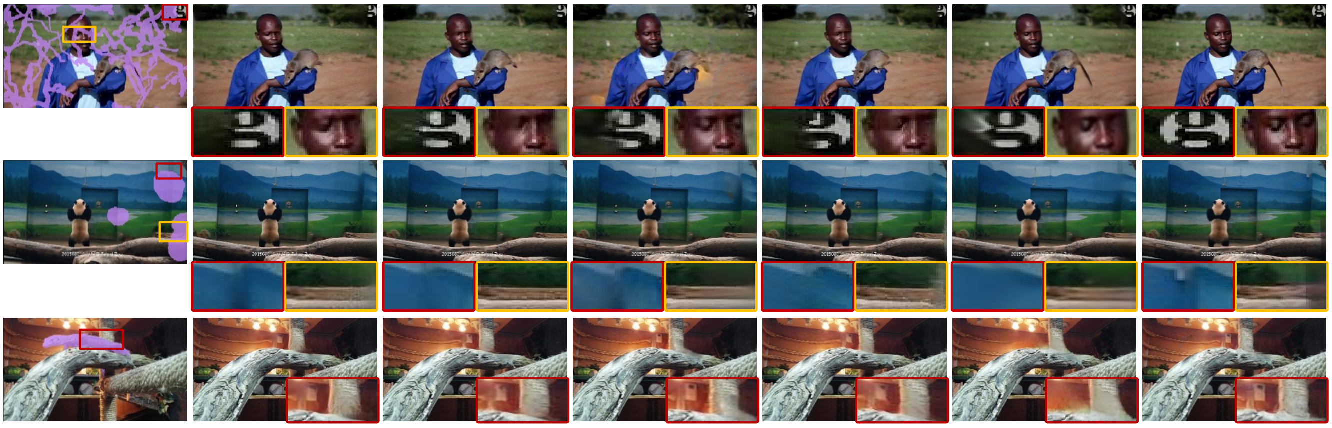

Figure 4: Qualitative results compared with FuseFormer [Liu *et al.*, 2021], E2FGVI [Li *et al.*, 2022], FGT [Zhang *et al.*, 2022b], ProPainter [Zhou *et al.*, 2023], and WaveFormer [Wu *et al.*, 2024] under three mask setting.

of the results, these baselines are fine-tuned using their released models and codes, and we report best results. Furthermore, PSNR [Haotian *et al.*, 2019], SSIM [Lin *et al.*, 2021], LPIPS [Zhang *et al.*, 2018], and E_{warp} [Lai *et al.*, 2018] are used to evaluate inpainting quality.

4.2 Experimental Results and Analysis

Quantitative Results. Tab. 1 shows quantitative results on Youtube-vos and DAVIS datasets under 256×256 resolution. As can be seen from Tab. 1, our method significantly outperforms all competitive baselines in four metrics. In particular, our method achieves 1.18%, 1.03%, 7.43% and 12.64% relative improvements on Youtube-vos dataset and 1.48%, 1.13%, 14.43%, and 8.57% relative improvements on DAVIS dataset regarding the PSNR, SSIM, E_{warp}, and LPIPS. These quantitative results validate that our proposed method can generate results with more visually realistic (PSNR, SSIM, and LPIPS), and more temporally consistent (E_{warp}).

Qualitative Results. In Fig. 4, we visually compare the qualitative results of our method with five baselines (FuseFormer [Liu *et al.*, 2021], E2FGVI [Li *et al.*, 2022], FGT [Zhang *et al.*, 2022b], ProPainter [Zhou *et al.*, 2023], and WaveFormer [Wu *et al.*, 2024]) under three different setting: (a) curve mask, (b) stationary mask, and (c) object mask. As can be seen, frames inpainted by our method have more realistic details which are significantly better than baselines. Our inpainted results not only have complete semantic structure, but also their details are more vivid and clear. Furthermore, to verify the effectiveness of "Drafting and Revision" framework, we compare the inpainting results of our method with ProPainter and WaveFormer on different frequencies in Fig. 5. As can be seen, the low-frequency semantics among these three methods exhibit negligible differences, whereas our method significantly outperforms ProPainter and WaveFormer in capturing high-frequency details. This indicates that our proposed "Drafting and Revision" framework can generate richer high-frequency details.

4.3 High Resolution Video inpainting

Recently, significant progress has been made by deep learning based video inpainting methods. However, due to the memory limitations of hardware devices, these methods can only complete videos with resolutions smaller than 1K. Naively using the"*down-sampling−inpainting−up-sampling*" technical pipeline to complete high-resolution video merely yield blurry results, which is disadvantageous in real-world applications. In our framework, Revision Network can be stacked more layers to handle high-resolution video. Tab. 2 shows the quantitative results at three different resolutions on Youtube-vos [Xu *et al.*, 2018] dataset, which were tested under a RTX 2080 Ti GPU. As shown in Tab. 2, as the video frame resolution increases, the baseline models suffer from GPU memory overflow. In contrast, our DRCN constructed by stacking three layers of Revision Network can effectively inpaint videos at a resolution of 2048×2048. Furthermore, Fig. 6 illustrates inpainted examples of our method at 2048×2048 resolution. It can be observed that DRCN generates missing contents with rich details for high-resolution videos.

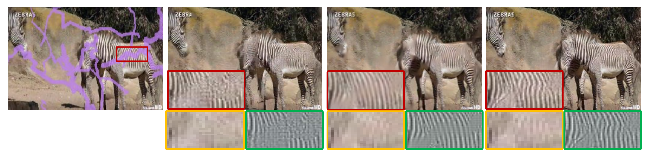

Figure 5: Visual comparisons at different frequencies, where the red, yellow and green boxes are the local slices, and its high-frequency and low-frequency components.

4.4 Ablation Study

Drafting Network. To demonstrate the effectiveness of Drafting Network, we replaced the Drafting Network in our framework with two baseline models (STTN [Zeng *et al.*, 2020], and E2FGVI [Li *et al.*, 2022]) and compared their results with our full model. As shown in Tab. 3, our model outperforms *STTN+ Revision* and *E2FGVI+Revision* on four metrics. These results demonstrate that the proposed Drafting Network is beneficial and necessary for completing the global semantic structure of the video frame.

Revision Network. In Fig. 7, we visually compared the inpainting results of *E2FGVI+Revision* and our full model. As observed from Fig. 7, the *E2FGVI+Revision* and our full

Methods	PSNR ↑	SSIM ↑	E_{warp} ↓	LPIPS ↓
	512×512 / 1024×1024 / 2048×2048 (**Resolutions**)			
VINet	26.363 / – / –	0.8770 / – / –	0.1642 / – / –	0.9926 / – / –
FGVC	24.411 / – / –	0.8453 / – / –	0.2120 / – / –	0.9812 / – / –
E2FGVI	32.558 / – / –	0.9293 / – / –	0.1221 / – / –	0.4031 / – / –
FGT	31.236 / – / –	0.9309 / – / –	0.1034 / – / –	0.3701 / – / –
STTN	28.176 / – / –	0.8486 / – / –	0.1587 / – / –	0.7074 / – / –
FuseFormer	29.613 / – / –	0.8754 / – / –	0.1479 / – / –	0.5605 / – / –
ProPainter	34.001 / – / –	0.9345 / – / –	0.0994 / – / –	0.3961 / – / –
WaveFormer	34.437 / – / –	0.9471 / – / –	0.1081 / – / –	0.2063 / – / –
CPVINet	31.629 / 32.117 / –	0.9265 / 0.9379 / –	0.1052 / 0.1031 / –	0.7211 / 0.6917 / –
Ours	**34.534 / 34.692 / 34.701**	**0.9537 / 0.9514 / 0.9634**	**0.0825 / 0.0964 / 0.1147**	**0.1860 / 0.1725 / 0.1767**

Table 2: Quantitative results of high resolution video on Youtube-vos dataset under a RTX 2080 Ti GPU. Note that certain models cause Out-Of-Memory (OOM) error when tested on 1K or 2K videos, thus the corresponding cells are empty, denoted as "–".

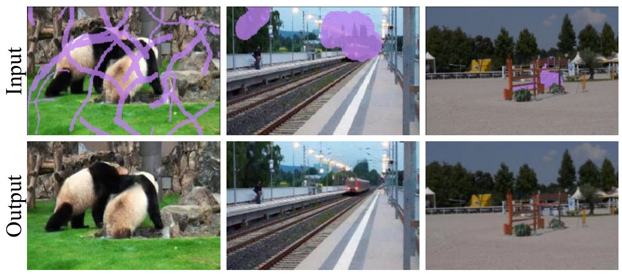

Figure 6: Inpainted results of our method at 2048 × 2048 resolution.

Modules	PSNR↑	SSIM↑	E_{warp} ↓	LPIPS↓
STTN	28.993	0.8761	0.1523	0.6965
STTN+Revision	30.258	0.8901	0.1485	0.6648
E2FGVI	30.064	0.9004	0.1490	0.5321
E2FGVI+Revision	31.416	0.9171	0.1402	0.5165
Drafting	31.291	0.9237	0.1423	0.3918
Full model	**33.658**	**0.9532**	**0.1096**	**0.2565**

Table 3: Ablation Study of Drafting Network and Revision Network.

model generate more reasonable and detailed missing contents than the E2FGVI [Li *et al.*, 2022] and Drafting Network. These results indicate that Revision Network has significant benefits for the final video inpainting results, which further verifies the necessity of the proposed "Drafting and Revision" framework in video inpainting task.

Alignment Manner. This section compares different temporal alignment manner. From Tab. 4, the following conclusions are confirmed: i) Performing temporal alignment can improve the performance ($2nd$-$6th$ rows). ii) The reference frame alignment effect of our alignment module ($6th$ row) outperforms traditional flow-based warping alignment methods ($2nd$ row) and traditional deformable convolution alignment methods ($3rd$ row). iii) Utilizing a lightweight estimator to calculate optical flow between frames does not significantly reduce the performance of the temporal alignment module ($5th$ row). iv) The strategy of aggregating the aligned reference feature $\widetilde{\boldsymbol{f}}_r$ and warped reference feature $\bar{\boldsymbol{f}}_r$ by Eq.(7) can obtain more a robust alignment feature ($6th$ row). v) Using the alignment loss $\mathcal{L}_a$ to train the temporal alignment module in a self-supervised manner can further improve the alignment performance of reference frame ($7th$ row).

Figure 7: Visual comparisons of inpainted results on E2FGVI, Drafting, E2FGVI+Revision, and our full model.

Alignment Manner	PSNR↑	SSIM↑	E_{warp} ↓	LPIPS↓
w/o align	27.696	0.8698	0.1567	0.5893
flow warping	32.108	0.9236	0.1491	0.4425
DCN	32.916	0.9279	0.1462	0.4110
flow + DCN	33.082	0.9303	0.1417	0.3971
ME + DCN	33.018	0.9296	0.1424	0.3979
ME + DCN + agg	33.363	0.9310	0.1406	0.3955
ME + DCN + agg + $\mathcal{L}_a$	**33.658**	**0.9532**	**0.1096**	**0.2565**

Table 4: Ablation study of diverse temporal alignment.

5 Conclusion

This paper propose a novel Drafting-and-Revision Completion Network (DRCN) for video inpainting, which contains two main sub-networks, *i.e.*, Drafting Network and Revision Network, where the former learns to complete the semantic information at low resolution, and the latter aims to generate the detailed information at high resolution. Such a design can flexibly supervise the inpainting of high-frequency and low-frequency component separately to effectively mitigate the inter-frequency conflicts. Furthermore, our DRCN can provide a feasible solution for high resolution video inpainting by stacking the Revision Network in a pyramid manner. Comprehensive experiments demonstrate the effectiveness of our model in both quantitative and qualitative evaluations.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (U2336212), the Natural Science Foundation of Zhejiang Province (LDT23F02023F02) and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

[Chan *et al.*, 2022] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proc. CVPR*, pages 5962–5971, 2022.

[Chang *et al.*, 2019] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proc. ICCV*, pages 9066–9075, 2019.

[Chen *et al.*, 2025] Kerui Chen, Zhiliang Wu, Wenjin Hou, Kun Li, Hehe Fan, and Yi Yang. Prompt-aware controllable shadow removal. *arXiv preprint arXiv:2501.15043*, 2025.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. ICCV*, pages 764–773, 2017.

[Deng *et al.*, 2019] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *Proc. ICCV*, pages 3076–3085, 2019.

[Gao *et al.*, 2020] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Proc. ECCV*, pages 713–729, 2020.

[Haotian *et al.*, 2019] Zhang Haotian, Mai Long, Wang Hailin, JinZha ando wen, and Ning Xu; John Collomosse. An internal learning approach to video inpainting. In *Proc. ICCV*, pages 2720–2729, 2019.

[Kim *et al.*, 2019] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep blind video decaptioning by temporal aggregation and recurrence. In *Proc. CVPR*, pages 4263–4272, 2019.

[Lai *et al.*, 2018] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proc. ECCV*, pages 179–195, 2018.

[Lee *et al.*, 2019] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proc. ICCV*, pages 4413–4421, 2019.

[Li *et al.*, 2019] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. *Proc. NeurIPS*, pages 1–13, 2019.

[Li *et al.*, 2020] 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 *Proc. ECCV*, pages 728–743, 2020.

[Li *et al.*, 2022] 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 *Proc. CVPR*, pages 17562–17571, 2022.

[Li *et al.*, 2023] Kun Li, Dan Guo, and Meng Wang. Vigt: proposal-free video grounding with a learnable token in the transformer. *SCI*, 66(10):202102, 2023.

[Li *et al.*, 2024] Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. Mmad: Multi-label micro-action detection in videos. *arXiv preprint arXiv:2407.05311*, 2024.

[Li *et al.*, 2025a] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. Prototypical calibrating ambiguous samples for micro-action recognition. In *Proc. AAAI*, pages 4815–4823, 2025.

[Li *et al.*, 2025b] Kun Li, Xinge Peng, Dan Guo, Xun Yang, and Meng Wang. Repetitive action counting with hybrid temporal relation modeling. *IEEE TMM*, 2025.

[Liang *et al.*, 2022] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. In *Proc. NeurIPS*, pages 378–393, 2022.

[Lin *et al.*, 2021] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proc. CVPR*, pages 5141–5150, 2021.

[Liu *et al.*, 2021] 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 *Proc. ICCV*, pages 14040–14049, 2021.

[Liu *et al.*, 2024a] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proc. CVPR*, pages 8038–8047, 2024.

[Liu *et al.*, 2024b] Pengyu Liu, Fei Wang, Kun Li, Guoliang Chen, Yanyan Wei, Shengeng Tang, Zhiliang Wu, and Dan Guo. Micro-gesture online recognition using learnable query points. *arXiv preprint arXiv:2407.04490*, 2024.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. CVPR*, pages 2536–2544, 2016.

[Perazzi *et al.*, 2016] 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 *Proc. CVPR*, pages 724–732, 2016.

[Ren *et al.*, 2022] Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *Proc. CVPR*, pages 3511–3520, 2022.

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, pages 99–121, 2000.

[Sun *et al.*, 2018] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proc. CVPR*, pages 8934–8943, 2018.

[Tian *et al.*, 2020] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proc. CVPR*, pages 3360–3369, 2020.

[Tian *et al.*, 2021] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. Coarse-to-fine cnn for image super-resolution. *TMM*, pages 1489–1502, 2021.

[Wang *et al.*, 2019] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proc. CVPR Workshops*, 2019.

[Wang *et al.*, 2023] Jianan Wang, Hanyu Xuan, and Zhiliang Wu. Text-video completion networks with motion compensation and attention aggregation. In *Proc. PRCV*, page 224–236, 2023.

[Wang *et al.*, 2024] Jianan Wang, Zhiliang Wu, Hanyu Xuan, and Yan Yan. Text-video completion networks with motion compensation and attention aggregation. In *Proc. ICASSP*, pages 2990–2994, 2024.

[Wang *et al.*, 2025] Fei Wang, Kun Li, Yiqi Nie, Zhangling Duan, Peng Zou, Zhiliang Wu, Yuwei Wang, and Yanyan Wei. Exploiting ensemble learning for cross-view isolated sign language recognition. *arXiv preprint arXiv:2502.02196*, 2025.

[Wu *et al.*, 2021] Zhiliang Wu, Kang Zhang, Hanyu Xuan, Jian Yang, and Yan Yan. DAPC-Net: Deformable alignment and pyramid context completion networks for video inpainting. *SPL*, pages 1145–1149, 2021.

[Wu *et al.*, 2023a] Zhiliang Wu, Changchang Sun, Hanyu Xuan, and Yan Yan. Deep stereo video inpainting. In *Proc. CVPR*, pages 5693–5702, 2023.

[Wu *et al.*, 2023b] Zhiliang Wu, Changchang Sun, Hanyu Xuan, Kang Zhang, and Yan Yan. Divide-and-conquer completion network for video inpainting. *TCSVT*, pages 2753–2766, 2023.

[Wu *et al.*, 2023c] Zhiliang Wu, Hanyu Xuan, Changchang Sun, Weili Guan, Kang Zhang, and Yan Yan. Semi-supervised video inpainting with cycle consistency constraints. In *Proc. CVPR*, pages 22586–22595, 2023.

[Wu *et al.*, 2023d] Zhiliang Wu, Kang Zhang, Changchang Sun, Hanyu Xuan, and Yan Yan. Flow-guided deformable alignment network with self-supervision for video inpainting. In *Proc. ICASSP*, pages 1–5, 2023.

[Wu *et al.*, 2024] Zhiliang Wu, Changchang Sun, Hanyu Xuan, Gaowen Liu, and Yan Yan. Waveformer: Wavelet transformer for noise-robust video inpainting. In *Proc. AAAI*, pages 6180–6188, 2024.

[Xiao *et al.*, 2023] Yi Xiao, Qiangqiang Yuan, Qiang Zhang, and Liangpei Zhang. Deep blind super-resolution for satellite video. *TGRS*, pages 1–16, 2023.

[Xu *et al.*, 2018] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proc. ECCV*, pages 603–619, 2018.

[Xu *et al.*, 2019] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proc. CVPR*, pages 3723–3732, 2019.

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

[Yu *et al.*, 2021] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proc. ICCV*, pages 14094–14103, 2021.

[Yu *et al.*, 2023] Yongsheng Yu, Heng Fan, and Libo Zhang. Deficiency-aware masked transformer for video inpainting. *arXiv preprint arXiv:2307.08629*, 2023.

[Zeng *et al.*, 2020] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *Proc. ECCV*, pages 3723–3732, 2020.

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

[Zhang *et al.*, 2022a] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *Proc. ECCV*, pages 581–596, 2022.

[Zhang *et al.*, 2022b] Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *Proc. ECCV*, pages 74–90, 2022.

[Zhang *et al.*, 2022c] Kaidong Zhang, Jingjing Fu, and Dong Liu. Inertia-guided flow completion and style fusion for video inpainting. In *Proc. CVPR*, pages 5982–5991, 2022.

[Zhang *et al.*, 2023] Yanni Zhang, Zhiliang Wu, and Yan Yan. Pfta-net: Progressive feature alignment and temporal attention fusion networks for video inpainting. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 191–195, 2023.

[Zhang *et al.*, 2024] Kaidong Zhang, Jialun Peng, Jingjing Fu, and Dong Liu. Exploiting optical flow guidance for transformer-based video inpainting. *TPAMI*, pages 1–16, 2024.

[Zhou *et al.*, 2023] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proc. ICCV*, pages 10477–10486, 2023.

[Zhuang *et al.*, 2024] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *Proc. ECCV*, pages 195–211, 2024.