WaveFormer: Wavelet Transformer for Noise-Robust Video Inpainting

Zhiliang Wu[1], Changchang Sun[2], Hanyu Xuan[3]*, Gaowen Liu[4], Yan Yan[2]

[1] CCAI, Zhejiang University, China
[2] Department of Computer Science, Illinois Institute of Technology, USA
[3] School of Big Data and Statistics, Anhui University, China
[4] Cisco Research, USA

Abstract

Video inpainting aims to fill in the missing regions of the video frames with plausible content. Benefiting from the outstanding long-range modeling capacity, the transformer-based models have achieved unprecedented performance regarding inpainting quality. Essentially, coherent contents from all the frames along both spatial and temporal dimensions are concerned by a patch-wise attention module, and then the missing contents are generated based on the attention-weighted summation. In this way, attention retrieval accuracy has become the main bottleneck to improve the video inpainting performance, where the factors affecting attention calculation should be explored to maximize the advantages of transformer. Towards this end, in this paper, we theoretically certificate that noise is the culprit that entangles the process of attention calculation. Meanwhile, we propose a novel wavelet transformer network with noise robustness for video inpainting, named WaveFormer. Unlike existing transformer-based methods that utilize the whole embeddings to calculate the attention, our WaveFormer first separates the noise existing in the embedding into high-frequency components by introducing the Discrete Wavelet Transform (DWT), and then adopts clean low-frequency components to calculate the attention. In this way, the impact of noise on attention computation can be greatly mitigated and the missing content regarding different frequencies can be generated by sharing the calculated attention. Extensive experiments validate the superior performance of our method over state-of-the-art baselines both qualitatively and quantitatively.

Introduction

Video inpainting which aims to fill missing regions of videos with plausible contents is a fundamental yet challenging task in the computer vision field. It has great value in many practical applications, such as scratch restoration (Chang et al. 2019), undesired object removal (Seoung et al. 2019) and autonomous driving (Liao et al. 2020). Unlike image inpainting (Somani et al. 2023; Shukla et al. 2023; Bar et al. 2022) that usually focuses on the spatial dimension, video inpainting pays more attention to exploiting the temporal information. Therefore, naively extending the image inpainting algorithm on individual video frame will neglect the inter-frame motion continuity, resulting in flicker artifacts (Chang et al. 2019; Wu et al. 2023a).

Recently, several deep learning-based video inpainting methods (Gao et al. 2020; Ji et al. 2022; Lee et al. 2019; Li et al. 2022; Wu et al. 2021; Zeng et al. 2019; Liu et al. 2020) have been proposed and achieved great progress in terms of the quality and speed. However, due to the limited receptive field along the temporal domain, these methods still suffer from limitations of blurry and misplacement artifacts in the completed video (Ren et al. 2022; Wu et al. 2023b). To address these issues, the state-of-the-art methods (Cai et al. 2022; Lee et al. 2019; Li et al. 2020; Liu et al. 2021; Ren et al. 2022; Seoung et al. 2019; Wu et al. 2023c; Zhang, Wu, and Yan 2023) resort to the attention mechanism to explore the long-term correspondences between frames. In this way, the available content at distant frames can also be globally propagated into the missing regions. Notably, the representative technique transformer (Cai et al. 2022; Liu et al. 2021; Ren et al. 2022; Zeng, Fu, and Chao 2020; Cai et al. 2022; Zhang, Fu, and Liu 2022) has gained increasing attention from researchers of video inpainting field due to its remarkable advantage of long-range modeling capacity. Typically, these transformer-based methods first search coherent contents from all the frames along both spatial and temporal dimensions by a patch-wise attention mechanism, and then utilize the attention-weighted summation to generate the missing contents. It means that the attention retrieval accuracy has become the main bottleneck limiting the inpainting performance. Inaccurate attention retrieval will ignore relevant content that is essential in video inpainting and introduce more irrelevant content in the missing regions, resulting in generating blurry or compromised contents (Zhang et al. 2023; Zhang, Fu, and Liu 2022).

In fact, due to the limitations of transmission media and recording equipment, digital images and videos will inevitably be polluted by noise during the transmission and recording process (Geng et al. 2022). Correspondingly, the learned embeddings always contain noise. Therefore, to improve the performance of video inpainting, it is promising and necessary to explore the impact of noise on attention computation. For this purpose, we theoretically certificate that noise-contained inputs are disadvantageous to transformers' attention calculation. Then, to address above disadvantages caused by ubiquitous noise in video inpainting, we

*Corresponding author.

propose a novel wavelet transformer network by introducing the Discrete Wavelet Transform (DWT) (Mallat 1989), dubbed as WaveFormer.

Concretely, unlike existing transformer-based video inpainting methods that utilize the whole embedding to calculate attention (Cai et al. 2022; Liu et al. 2021; Ren et al. 2022; Zhang, Fu, and Liu 2022), our WaveFormer first adopts DWT to decompose the embedding used for the attention calculation into low-frequency and high-frequency components. By doing this, the noise existing in the embedding can be explicitly separated into the high-frequency components, making the low-frequency ones contain relatively clean basic features. In this way, the calculation of attention weight is only based on the low-frequency components and the missing content regarding different frequencies can be generated by sharing such attentions. Finally, the completed low-frequency and high-frequency components are aggregated to yield the final inpainting result through Inverse Discrete Wavelet Transform (IDWT).

Substantial experiments show that our WaveFormer outperforms the state-of-the-arts by a significant margin in terms of PSNR and E_{warp} (flow warping error) with relative improvements of 7.45% and 9.48%, respectively. Moreover, thanks to the robustness to noise, our method is able to fill missing regions using the visually-plausible and spatial-temporal coherent contents with fine-grained details. To sum up, our contributions are summarized as follows:

- We theoretically demonstrate that noise always cause inferior effect when calculating attention. To the best of our knowledge, this is the first attempt to explore the factors that affect the transformers' attention calculation in the video inpainting.
- We propose a novel WaveFormer by introducing the DWT. It can calculate the attention on low-frequency components and share it with the high-frequency components, greatly mitigating the impact of noise on the attention calculation.
- Experiments on two benchmark datasets, including Youtube-vos (Xu et al. 2018) and DAVIS (Perazzi et al. 2016), demonstrate the superiority of our proposed method in both quantitative and qualitative views.

Related Work

Video Inpainting

With the rapid development of deep learning (Shang et al. 2023; Gu et al. 2023; Shang et al. 2022), several deep learning-based video inpainting methods have been proposed recently. For instance, Wang et al. (Wang et al. 2019), Kim et al. (Kim et al. 2019), and Chang et al. (Chang et al. 2019) employed the 3D temporal convolution and directly aggregate the temporal information of neighbor frames to reconstruct the missing contents. However, compared with 2D CNN, 3D CNN has relatively higher computational complexities, limiting the application of these methods in the real scenarios (Wu et al. 2023b; Ji et al. 2022; Liu, Li, and Zhu 2022). To alleviate this issue, treating the video inpainting as a pixel propagation problem has been explored by some works (Gao et al. 2020; Kang, Oh, and Kim 2022; Ke, Tai, and Tang 2021; Li et al. 2022; Xu et al. 2019; Zou et al. 2021). In particular, they first exploit a deep flow completion network to restore the flow sequence. Such a restored flow sequence is used to guide the relevant pixels of neighboring frames to fill in the missing regions. Overall, although these methods have shown promising results, they fail to capture the visible contents of long-distance frames, resulting in poor inpainting performance in the scene with large objects or slowly moving objects.

To effectively model the long-distance correspondence, recent methods (Cai et al. 2022; Li et al. 2020; Ren et al. 2022; Seoung et al. 2019; Srinivasan et al. 2021) introduced the attention module to retrieve information from neighboring frames and adopted weighted summing operation to generate missing contents. Among these methods, benefiting from the advantages of long-range feature capture capacity, transformer has shed light to the video inpainting community. For example, Zeng et al. (Zeng, Fu, and Chao 2020) proposed the first transformer model for video inpainting by designing a multi-layer multi-head transformer. To improve the edge details of missing contents, Liu et al. (Liu et al. 2021) devised a new transformer model by introducing soft split and soft composition operations. In addition, Ren et al. (Ren et al. 2022) developed a novel Discrete Latent Transformer (DLFormer) by formulating video inpainting task into the discrete latent space. Meanwhile, Zhang (Zhang, Fu, and Liu 2022) leveraged the motion discrepancy exposed by optical flows to instruct the attention retrieval in the transformer for high-fidelity video inpainting. At the same time, Cai (Cai et al. 2022) designed a new Deformed Vision Transformer (DeViT) with emphasis on better patch-wise alignment and matching in video inpainting.

It is worth noting that these transformer-based video inpainting methods ignore the impact of noise on attention calculation, which inevitably leads to inaccurate attention retrieval. In our work, we expect to explore the mechanism of noise works in the attention calculation and propose a novel wavelet transformer network with noise robustness to improve the accuracy of attention retrieval.

Discrete Wavelet Transform (DWT)

Thanks to the powerful time-frequency analysis capability of DWT, more and more researchers expect to combine it with deep learning to solve various computer vision tasks. For example, Liu et al. (Liu et al. 2018) presented a novel multi-level wavelet CNN to enlarge the receptive field for a better trade-off between efficiency and restoration performance. To preserve the original image details while reducing computational cost in self-attention learning, Yao et al. (Yao et al. 2022) formulated a invertible downsampling for wavelet transforms. Yu et al. (Yu et al. 2021) proposed a wavelet-based inpainting network that can separately fills the missing regions of each frequency band. These works show that combining wavelets and CNNs is promising. However, to the best of our knowledge, the potential of using wavelets to mitigate the influence of noise on the attention calculation of transformer has not been well validated, which is the major concern of this paper.

Motivation

In this paper, we argue that noise is disadvantageous to the transformers' attention calculation, which greatly limits the performance of video inpainting. Using the noise-contained embeddings to calculate the attention will disregard the contents related with the missing regions and increases unrelated contents filled into the missing regions during video completion, leading to blurred or compromised missing contents and hence suffer from the inferior inpainting results.

Theorem: Given n noise-contained features $\boldsymbol{f}_i$, whose dimension is $h \times w \times c$ and value range from 0 to 1. Formally, $\boldsymbol{f}_i$ can be denoted as the summation of the clean feature $\boldsymbol{e}_i \in [0,1]^{h\times w\times c}$ and the noise $\boldsymbol{o}_i \in [0,1]^{h\times w\times c}$ (Cheng et al. 2021; Jia, Wong, and Zeng 2021; Pang et al. 2021), *i.e.*, $\boldsymbol{f}_i = \boldsymbol{e}_i + \boldsymbol{o}_i$. Let $r^f_{i,j}$ stands for the attention between noise-contained features $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, and $r^e_{i,j}$ denotes the attention between $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, which can be obtained as follows,

$$r^f_{i,j} = \frac{\exp(s^f_{i,j})}{\sum_{t=1}^n \exp(s^f_{i,t})},\ \ r^e_{i,j} = \frac{\exp(s^e_{i,j})}{\sum_{t=1}^n \exp(s^e_{i,t})}, \tag{1}$$

where $s^f_{i,j} = \frac{\boldsymbol{f}_i \boldsymbol{f}_j^T}{\sqrt{h\times w\times c}}$, $s^e_{i,j} = \frac{\boldsymbol{e}_i \cdot \boldsymbol{e}_j^T}{\sqrt{h\times w\times c}}$. Essentially, the value of $r_{i,j}$ represents the correlation extent between two features. The correlation reaches maximum when $r_{i,j} = 1$, representing i-th feature is completely related to the j-th feature, and vice versa.

Based on above theory regarding attention, the following theoretical statements hold:

- if $r^e_{i,j} \to 0$, then $r^e_{i,j} < r^f_{i,j}$, *i.e.*, the noise increases the attention between unrelated contents;
- if $r^e_{i,j} \to 1$, then $r^e_{i,j} > r^f_{i,j}$, *i.e.*, the noise decreases the attention between related contents.

Proof: According to the definition of the $r^f_{i,j}$, we have:

$$r^f_{i,j} = \frac{\exp(s^f_{i,j})}{\sum_{t=1}^n \exp(s^f_{i,t})} = \frac{\exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_j^T)}{\sum_{t=1}^n \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_t^T)}. \tag{2}$$

Simple algebra computations enable us to have,

$$\frac{r^e_{i,j}}{r^f_{i,j}} = r^e_{i,j}\left(\frac{\sum_{t=1,\neq j}^n \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_t^T)}{\exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_j^T)} + 1\right). \tag{3}$$

Besides, as $\exp(x)$ is a monotonically increasing function and its value ranges from 0 to 1, we have,

$$\begin{aligned} & 1 \le \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_t^T) \le e, \\ \Rightarrow\ & n-1 \le \sum_{t=1,\neq j}^{n} \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_t^T) \le (n-1)e, \\ \Rightarrow\ & \frac{n-1}{e} \le \frac{\sum_{t=1,\neq j}^n \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_t^T)}{\exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_j^T)} \le (n-1)e. \end{aligned} \tag{4}$$

Since n is a finite real number, $\frac{n-1}{e}$ and $(n-1)e$ are both finite real numbers. Considering the convenience of the expression, we denote $\sum_{t=1,\neq j}^n \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_t^T)/\exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_j^T)$ revealed in Eq.(3) and (4) as $\mathcal{F}_{ijt}$. For each statement of above theorem, we can prove it as follows,

1) if $r^e_{i,j} \to 0$, we have,

$$r^e_{i,j}\left(\mathcal{F}_{ijt}+1\right) \to 0,\ \Rightarrow\ \frac{r^e_{i,j}}{r^f_{i,j}} \to 0{<}1,\ \Rightarrow\ r^e_{i,j}{<}r^f_{i,j}, \tag{5}$$

2) if $r^e_{i,j} \to 1$, we have,

$$r^e_{i,j}\left(\mathcal{F}_{ijt}+1\right) {>} 1,\ \Rightarrow\ \frac{r^e_{i,j}}{r^f_{i,j}} {>} 1,\ \Rightarrow\ r^e_{i,j}{>}r^f_{i,j}. \tag{6}$$

Methodology

Formulation and Overview

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$ be a corrupted video sequence consisting of T frames with height H and width W. The corresponding frame-wise masks are denoted as $\boldsymbol{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_T\}$. For each mask $\boldsymbol{m}_i$, "0" indicates that corresponding pixel is valid, and "1" denotes that the pixel is missing or corrupted. The goal of video inpainting is to generate an inpainted video sequence $\widehat{\boldsymbol{Y}} = \{\widehat{\boldsymbol{y}}_1, \widehat{\boldsymbol{y}}_2, \cdots, \widehat{\boldsymbol{y}}_T\}$, which are spatially and temporally consistent with the original video sequence $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_T\}$.

Based on the fact that the contents of missing regions in one frame may exist in neighboring frames, existing transformer-based methods (Cai et al. 2022; Liu et al. 2021; Ren et al. 2022; Zhang, Fu, and Liu 2022; Yu, Fan, and Zhang 2023; Zhang et al. 2023) usually formulate the video inpainting task as a "multi-to-multi" conditional distribution prediction problem as follows,

$$p(\widehat{\boldsymbol{Y}}|\boldsymbol{X}) = \prod_{t=1}^{T} p(\widehat{\boldsymbol{Y}}^{t+n}_{t-n} | \boldsymbol{X}^{t+n}_{t-n}, \boldsymbol{M}^{t+n}_{t-n}), \tag{7}$$

where $\boldsymbol{X}^{t+n}_{t-n} = \{\boldsymbol{x}_{t-n}, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_{t+n}\}$ stands for a short clip of neighboring frames with a center moment t and a temporal radius n, $\boldsymbol{M}^{t+n}_{t-n}$ denotes the mask clip regarding $\boldsymbol{X}^{t+n}_{t-n}$. In practice, these transformer-based methods usually generate the missing contents by aggregating coherent contents, which are searched by patch-based attention module from all the frames along both spatial and temporal dimensions. Therefore, the attention retrieval accuracy is an critical factor affecting the final inpainting performance.

Inevitably, digital images and videos are polluted by noise during the transmission and recording process (Geng et al. 2022), resulting in the learned embeddings always contain noise. In Sect., we have also theoretically confirmed that noise can also have an adverse effect on transformer-based video inpainting. For this purpose, we propose a novel wavelet transformer network with noise robustness to mitigate this adverse effect. As shown in Fig.1, the proposed WaveFormer mainly consists of three parts: a frame-level encoder, wavelet spatial-temporal transformer and a frame-level decoder. Specifically, the frame-level encoder is built by stacking multiple convolutional layers and residual blocks with ReLUs as activation functions, aiming to extract deep features from low-level pixels of each frame. Similarly, the frame-level decoder is designed to decode inpainted features into frames.

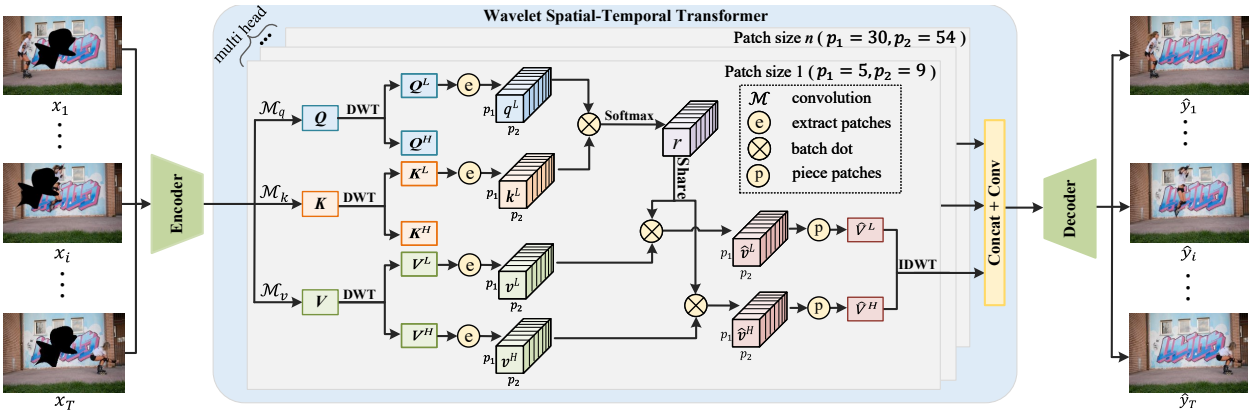


Figure 1: Illustration of the proposed WaveFormer, consisting of *1)* a frame-level encoder, *2)* the wavelet spatial-temporal transformer and *3)* a frame-level decoder. Instead of using queries ($\boldsymbol{Q}$) and keys ($\boldsymbol{K}$) to directly calculate attention in existing transformer-based methods, our WaveFormer employs Discrete Wavelet Transform (DWT) to separate the embedding into high-frequency and low-frequency components. These separated low-frequency components are relatively clean, which are used to calculate attention for video inpainting. In this way, the impact of noise on the attention weight is greatly mitigated.

Wavelet Spatial-Temporal Transformer

As the core component of our WaveFormer, wavelet spatial-temporal transformer is designed to search coherent contents from all the input frames with the aim of learning spatial-temporal transformations for all missing regions in the wavelet domain of the deep encoding space. Specifically, in the process of attention calculation, we introduce DWT to separate noise into high-frequency components, and then use low-frequency components to calculate attention. Finally, the calculated attention is shared with high-frequency components to generate the missing content with different frequencies. In this way, the impact of noise on attention computation can be greatly mitigated. Essentially, our wavelet spatial-temporal transformer also follows the general pipeline of transformer design, namely embedding, matching, and aggregating. We will introduce more details of each step one by one as below.

Embedding: Embedding aims to map deep features into key and memory, so as to establish deep correspondences for each region in different semantic spaces (Ren et al. 2022). Let $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_T\}$ denote the deep features encoded by the frame-level encoder, where $\boldsymbol{f}_i \in \mathbb{R}^{h \times w \times c}$. The three basic elements of the attention mechanism are extracted by the 1×1 convolution, including $\boldsymbol{Q}$(query), $\boldsymbol{K}$(key), and $\boldsymbol{V}$(value):

$$\boldsymbol{Q}_i, (\boldsymbol{K}_i, \boldsymbol{V}_i) = \mathcal{M}_q(\boldsymbol{f}_i), (\mathcal{M}_k(\boldsymbol{f}_i), \mathcal{M}_v(\boldsymbol{f}_i)), \tag{8}$$

where $1 \leq i \leq T$. $\mathcal{M}_q(\cdot)$, $\mathcal{M}_k(\cdot)$ and $\mathcal{M}_v(\cdot)$ denote the 1×1 2D convolution.

Matching: Having obtained these three basic elements, the coherent contents are searched by calculating the similarity between patches. Specifically, we first decompose $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$ and $\boldsymbol{V}_i$ into corresponding low-frequency components and high-frequency components by DWT, individually,

$$\boldsymbol{Q}_i^L, \boldsymbol{Q}_i^H;\ \boldsymbol{K}_i^L, \boldsymbol{K}_i^H;\ \boldsymbol{V}_i^L, \boldsymbol{V}_i^H = \mathcal{DWT}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i), \tag{9}$$

where $\boldsymbol{Q}_i^L, \boldsymbol{K}_i^L, \boldsymbol{V}_i^L \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times c}$ denote the low-frequency components corresponding to $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$ and $\boldsymbol{V}_i$, mainly recording principal information including the basic structures. Similarly, $\boldsymbol{Q}_i^H, \boldsymbol{K}_i^H, \boldsymbol{V}_i^H \in \mathbb{R}^{3 \times \frac{h}{2} \times \frac{w}{2} \times c}$ denote the high-frequency components in horizontal, vertical and diagonal directions, containing a very large proportion of data noise.

After obtaining the low-frequency components $\boldsymbol{Q}_i^L$ and $\boldsymbol{K}_i^L$, we extract spatial patches of shape $p_1 \times p_2 \times c$ from $\boldsymbol{Q}_i^L$ and $\boldsymbol{K}_i^L$ of each frame, denoted as $\boldsymbol{q}_i^L$ and $\boldsymbol{k}_i^L$. Then, the patch-wise similarities can be calculated by matrix multiplication, denoted as

$$s_{i,j} = \frac{\boldsymbol{q}_i^L \cdot (\boldsymbol{k}_j^L)^T}{\sqrt{p_1 \times p_2 \times c}}, \tag{10}$$

where $1 \leq i, j \leq N$ and $N = T \times \frac{h}{p_1} \times \frac{w}{p_2}$. A softmax function is introduced to obtain the attention weights of all patches,

$$r_{i,j} = \begin{cases} \exp(s_{i,j}) / \sum_{t=1}^{N} \exp(s_{i,t}), & \boldsymbol{q}_i^L \in \Omega, \\ 0, & \boldsymbol{q}_i^L \in \bar{\Omega}, \end{cases} \tag{11}$$

where Ω and $\bar{\Omega}$ denote visible regions and missing regions, respectively. Naturally, we only borrow features from visible regions to fill missing regions.

Aggregating: After modeling the deep correspondences of all spatial patches, we share the calculated attention on the low-frequency components with the high-frequency components. The output of the query for the low-frequency and high-frequency components of each patch can be obtained by the attention-weighted summation of the values of related patches, separately,

$$\hat{\boldsymbol{v}}_i^L = \sum_{j=1}^{N} r_{i,j} \boldsymbol{v}_j^L, \quad \hat{\boldsymbol{v}}_i^H = \sum_{j=1}^{N} r_{i,j} \boldsymbol{v}_j^H, \tag{12}$$

where $\boldsymbol{v}_j^L$ and $\boldsymbol{v}_j^H$ denote the value of the low-frequency and high-frequency components of the j-th patch, respectively.

Methods	YouTube-VOS (Xu et al. 2018)				DAVIS (Perazzi et al. 2016)			
	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓
TCCDS (Huang et al. 2016)	23.418	0.8119	0.3388	1.9372	28.146	0.8826	0.2409	1.0079
VINet (Kim et al. 2020)	26.174	0.8502	0.1694	1.0706	29.149	0.8965	0.1846	0.7262
DFVI (Xu et al. 2019)	28.672	0.8706	0.1479	0.6285	30.448	0.8961	0.1640	0.6857
FGVC (Gao et al. 2020)	24.244	0.8114	0.2484	1.5884	28.936	0.8852	0.2122	0.9598
CPVINet (Lee et al. 2019)	28.534	0.8798	0.1613	0.8126	30.234	0.8997	0.1892	0.6560
OPN (Seoung et al. 2019)	30.959	0.9142	0.1447	0.4145	32.281	0.9302	0.1661	0.3876
STTN (Zeng, Fu, and Chao 2020)	28.993	0.8761	0.1523	0.6965	28.891	0.8719	0.1844	0.8683
FuseFormer (Liu et al. 2021)	29.765	0.8876	0.1463	0.5481	29.627	0.8852	0.1767	0.6706
E2FGVI (Li et al. 2022)	30.064	0.9004	0.1490	0.5321	31.941	0.9188	0.4579	0.6344
FGT (Zhang, Fu, and Liu 2022)	30.811	0.9258	0.1308	0.4565	32.742	0.9272	0.1669	0.4240
WaveFormer	**33.264**	**0.9435**	**0.1184**	**0.2933**	**34.169**	**0.9475**	**0.1504**	**0.3137**

Table 1: Quantitative results of video inpainting on YouTube-VOS (Xu et al. 2018) and DAVIS (Perazzi et al. 2016) datasets.

We piece all patches together to acquire $\widehat{\boldsymbol{V}}_i^L \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c}$ and $\widehat{\boldsymbol{V}}_i^H \in \mathbb{R}^{3\times\frac{h}{2}\times\frac{w}{2}\times c}$, and then generate the completed feature $\widehat{\boldsymbol{f}}_i$ by IDWT:

$$\widehat{\boldsymbol{f}}_i = \mathcal{IDWT}(\widehat{\boldsymbol{V}}_i^L, \widehat{\boldsymbol{V}}_i^H). \tag{13}$$

Note that the proposed wavelet spatial-temporal transformer adopts a multi-head design, where different heads are employed to calculate the attention weights of patches with various sizes. In this way, the patches with large size can apply global features to complete semantic background, while the patches with small size can utilize local features to generate detailed texture, thereby achieving high-quality video inpainting. Furthermore, to fully exploit the power of the proposed transformer, our WaveFormer stacks multiple layers of the wavelet spatial-temporal transformer. Such a design can use the updated region features in a single feed-forward process to improve the results of attention to missing regions. The final inpainted frame $\widehat{\boldsymbol{y}}_i$ can be obtained by decoding $\widehat{\boldsymbol{f}}_i$ with the frame-level decoder.

Loss Function

The total loss of our WaveFormer consists of three terms, *i.e.*, the reconstruction term of the hole regions $\mathcal{L}_{hole}$ (Zeng, Fu, and Chao 2020), the reconstruction term of the valid regions $\mathcal{L}_{val}$ (Zeng, Fu, and Chao 2020) and the adversarial term $\mathcal{L}_{adv}$ by using Temporal PatchGAN (T-PatchGAN) (Chang et al. 2019) as a discriminator:

$$\mathcal{L} = \lambda_{hole}\mathcal{L}_{hole} + \lambda_{val}\mathcal{L}_{val} + \lambda_{adv}\mathcal{L}_{adv}, \tag{14}$$

where λ_{hole}, λ_{val} and λ_{adv} are the trade-off parameters. In real implementation, we empirically set these three parameters as 3, 5 and 0.01.

Experiments

Experimental Setting

Datasets and Evaluation Metrics. Two most commonly-used datasets are taken to verify the effectiveness of the proposed method, including Youtube-vos dataset (Xu et al. 2018) and DAVIS dataset (Perazzi et al. 2016). The former contains 3,471, 474 and 508 video clips in training, validation and test set, respectively. The latter is composed of 60 video clips for training and 90 video clips for testing. Following previous works, we report quantitative results by four metrics, including PSNR (Haotian et al. 2019), SSIM (Zhang et al. 2022), LPIPS (Zhang et al. 2018) and flow warping error E_{warp} (Lai et al. 2018).

Mask Settings. In the real world, the applications of video inpainting mainly include undesired object removal, scratch restoration, watermark removal, etc. To simulate these applications, we evaluate the model with the following three types of masks:

- Object mask: it is used to simulate applications like undesired object removal. Following FuseFormer (Liu et al. 2021), we employ the foreground object annotations in DAVIS dataset as the testing object masks, which have continuous motion and realistic appearance.
- Curve mask: it is composed of curves with continuous motion, which is exploited to simulate applications like scratch restoration. In our experiment, these curve masks are sampled from FVI dataset (Chang et al. 2019).
- Stationary mask: it has an arbitrary shapes but a relatively fixed position. The stationary mask is used to simulate applications such as watermark removal, and its generation process follows previous work (Chang et al. 2019; Zeng, Fu, and Chao 2020).

Experimental Results and Analysis

Quantitative Results. Quantitative results of video inpainting are reported on both YouTube-VOS and DAVIS. We select the most recent and the most competitive approaches as the baselines, including TCCDS, VINet, CPVINet, DFVI, FGVC, OPN, STTN, FuseFormer, E2FGVI and FGT. To ensure the comparability of experimental results, these baselines are fine-tuned several times based on their released codes, and report their best results in this section.

As shown in Tab.1, the PSNR, SSIM, E_{warp} and LPIPS of our model substantially surpass all previous state-of-the-art methods on YouTube-VOS and DAVIS. The superior results demonstrate that our WaveFormer can generate the videos with less distortion (PSNR and SSIM), more visually plausible content (LPIPS) and better spatial and temporal coherence (E_{warp}). Such a commendable performance verifies the superiority of the proposed method.

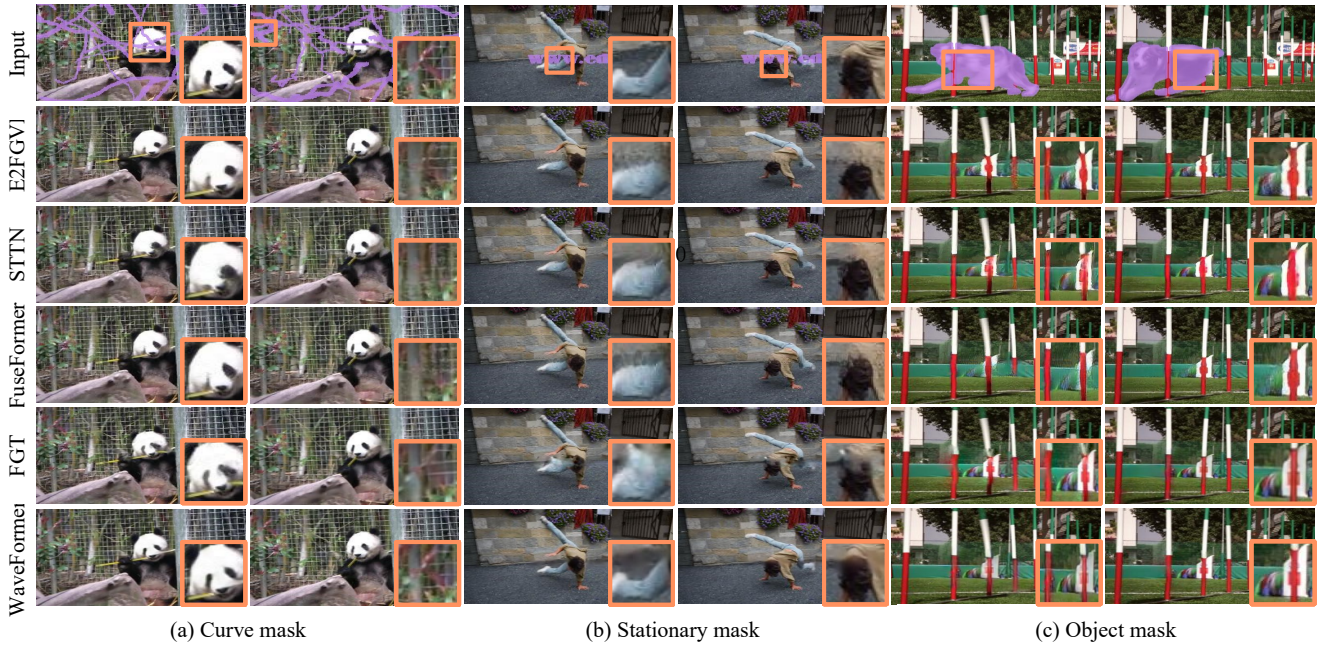


Figure 2: Qualitative results compared with E2FGVI (Li et al. 2022), STTN (Zeng, Fu, and Chao 2020), FuseFormer (Liu et al. 2021) ,and FGT (Zhang, Fu, and Liu 2022). Better viewed at zoom level 400%.

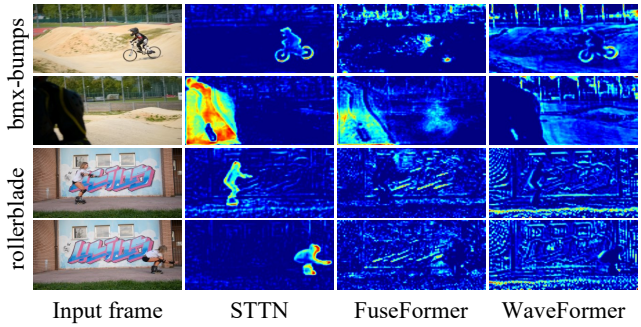


Figure 3: Comparison of the feature maps before feeding into transformer blocks between STTN (Zeng, Fu, and Chao 2020), FuseFormer (Liu et al. 2021) and our WaveFormer.

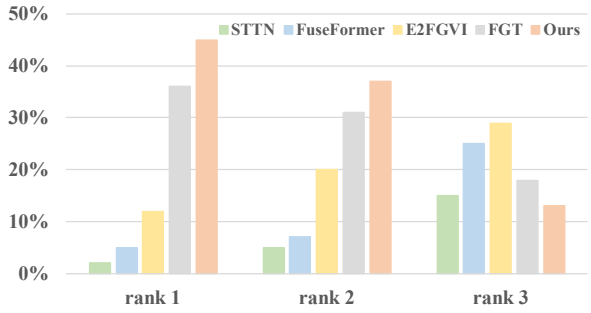
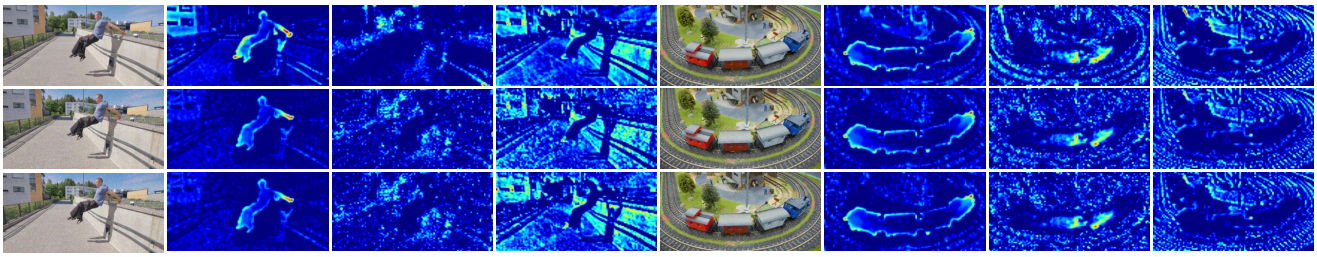


Figure 4: User study. "rank x" means the percentage of results from each model being chosen as the x-th best.

Qualitative Results. To visually inspect the visual results, we choose four competitive methods, including E2FGVI, STTN, FuseFormer and FGT, to conduct visual comparisons. Respectively, Fig.2 (a), Fig.2 (b) and Fig.2 (c) illustrates the scratch restoration case of curve masks, the watermark removal case of stationary masks and the object removal case of object masks. It can be observed that our WaveFormer generates the missing contents with more accurate structures and details than baselines in these three cases.

Furthermore, we also visualize the feature maps of STTN, FuseFormer and WaveFormer before extracting spatial patches for attention computation. As shown in the second example (rollerblade) of Fig. 3, the texture structure of text, windows and walls in the feature map generated by STTN is completely destroyed. Although the texture structure of text in the feature map produced by FuseFormer is retained, the texture structure of windows and walls has been totally broken by strong noise. Compared with these two most competitive approaches, our WaveFormer produces the feature map with a cleaner background and a more complete texture structure. It is easy to figure out the text, window and wall in our feature map. Such a distinct background texture leads to more accurate attention retrieval in the transformer block, thus naturally producing better visual quality. The above observations illustrate that noise accumulation destroys the texture structure used for attention retrieval, and our WaveFormer relieves this drawback to some extent. We believe that this is the reason why our WaveFormer has better inpainting performance.

User Study. In order to further make a comprehensive comparison, we conduct a user study of the inpainting results



(a) Parkour (b) Train

Figure 5: Visual comparison of the feature maps sourced from clean and noisy video frame, where the first, second and third rows are the clean frames, frame with Gaussian noise and frame with Salt & pepper noise, respectively. Best viewed in zoom.

of the five competitive approaches, including STTN, FuseFormer, E2FGVI, FGT and WaveFormer. We invited 20 volunteers to perform a questionnaire survey for 10 videos from the DAVIS dataset. In each inquiry, we asked volunteers to choose the video for which they think the inpainting results are best. To ensure the reliability of subjective evaluations, the inpainting results obtained by the five methods were scrambled at each interrogation, and each video can be played multiple times. The results of the user study are concluded in Fig. 4. As we can see, volunteers obviously favor our results compared with other competitors.

Ablation Study

Noise-robustness. Fig. 5 shows the feature maps with noisy frames as inputs in two representative example, where the first row reveals the feature map produced when using the clean frame from DAVIS dataset as inputs, and the next two rows display the feature map generated when using the frame added with Gaussian and Salt & pepper noise as inputs. As shown in Fig. 5, we can find that it is difficult for STTN and FuseFormer to suppress noise, while WaveFormer could suppress the noise and maintain the background structure during its inference. For example, in Fig. 5(a), the building structure in the two feature maps generated by STTN and WaveFormer is complete, when the clean parkour frame is fed. After the frame is superposed with Gaussian or salt & pepper noise, the feature map of STTN contains very strong noise, and the building structure vanishes, while the basic structure could still be observed from our WaveFormer. Similarly, in Fig. 5(b), the feature map of FuseFormer also contains strong noise, and the railway structure disappears, while WaveFormer can still observe the railway structure. such results indicate that WaveFormer is robust to different noises.

The Impact of Noise. To further verify the impact of noise on attention calculation, we use DWT to separate the noise from the embedding used for attention calculation in the STTN and FuseFormer, and compare them with its original versions. Here, the improved STTN and FuseFormer are labeled STTN_Wave and FuseFormer_Wave. As shown in Tab.2, STTN_Wave and FuseFormer_Wave are obviously superior to original STTN and FuseFormer in all evaluation metrics. These results demonstrate the effectiveness and necessity of noise removal in attention calculation.

Methods	PSNR↑	SSIM↑	$E_{warp}\downarrow$	LPIPS↓
STTN	28.993	0.8761	0.1523	0.6965
STTN_Wave	30.012	0.8917	0.1509	0.6631
FuseFormer	29.765	0.8876	0.1463	0.5481
FuseFormer_Wave	31.171	0.8995	0.1429	0.5236
w/o DWT	31.326	0.9259	0.1299	0.3471
Full model	33.264	0.9435	0.1184	0.2933

Table 2: Impact of noise on attention computation.

Methods	STTN	FuseFormer	E2FGVI	FGT	WaveFormer
FLOPs	477.91G	579.82G	442.18G	455.91G	349.71G
Time	0.22s	0.30s	0.26s	0.39s	0.18s

Table 3: Efficiency analysis.

Efficiency analysis. In addition, we compare the efficiency of WaveFormer with STTN, FuseFormer, E2FGVI and FGT by using FLOPs and inference time. Since the FLOPs in video inpainting are related to the simultaneous processing of the temporal size (number of frames), we set the temporal size to 20 following to previous works (Liu et al. 2021; Zeng, Fu, and Chao 2020; Zhang, Fu, and Liu 2022). And the runtime is measured on a single Titan RTX GPU. The compared results are shown in Tab. 3. The inference speed of the proposed method is the fastest, improving 0.04s over the optimal baseline—STTN. Besides, WaveFormer holds the lowest FLOPs in contrast to all other methods.

Conclusion

In this work, we theoretically proved that noise reduces the attention to relevant contents and increases the attention to irrelevant contents when generating the missing regions. Based on this fact, we propose a novel transformer network by introducing the DWT, named WaveFormer. Our WaveFormer uses DWT to separate the noise existing in the embedding into high-frequency components, and employs relatively clean low-frequency components to calculate attention weight, thereby mitigating the impact of noise on the calculation of attention weight to the greatest extent. Experiments demonstrate the superior performance of the proposed WaveFormer both quantitatively and qualitatively.

References

Bar, A.; Gandelsman, Y.; Darrell, T.; Globerson, A.; and Efros, A. 2022. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems (NIPS)*, volume 35, 25005–25017.

Cai, J.; Li, C.; Tao, X.; Yuan, C.; and Tai, Y.-W. 2022. DeViT: Deformed Vision Transformers in Video Inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia (ACMMM)*, 779–789.

Chang, Y.-L.; Liu, Z. Y.; Lee, K.-Y.; and Hsu, W. 2019. Free-form Video Inpainting with 3D Gated Convolution and Temporal PatchGAN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 9066–9075.

Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; and Liu, S. 2021. NBNet: Noise Basis Learning for Image Denoising With Subspace Projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4896–4906.

Gao, C.; Saraf, A.; Huang, J.-B.; and Kopf, J. 2020. Flow-edge Guided Video Completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 713–729.

Geng, M.; Meng, X.; Zhu, L.; Jiang, Z.; Gao, M.; Huang, Z.; Qiu, B.; Hu, Y.; Zhang, Y.; Ren, Q.; and Lu, Y. 2022. Triplet Cross-Fusion Learning for Unpaired Image Denoising in Optical Coherence Tomography. *IEEE Transactions on Medical Imaging (TMI)*, 41(11): 3357–3372.

Gu, B.; Yu, Y.; Fan, H.; and Zhang, L. 2023. Flow-Guided Diffusion for Video Inpainting. *arXiv preprint arXiv:2311.15368*.

Haotian, Z.; Long, M.; Hailin, W., JinZha ando wen; and Collomosse, N. X. J. 2019. An Internal Learning Approach to Video Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2720–2729.

Huang, J.; Kang, S. B.; Ahuja, N.; and Kopf, J. 2016. Temporally coherent completion of dynamic video. *ACM Transactions on Grapics (TOG)*, 35(6): 196.1–196.11.

Ji, Z.; Hou, J.; Su, Y.; Pang, Y.; and Li, X. 2022. G2LP-Net: Global to Local Progressive Video Inpainting Network. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33(3): 1082–1092.

Jia, F.; Wong, W. H.; and Zeng, T. 2021. DDUNet: Dense Dense U-Net With Applications in Image Denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 354–364.

Kang, J.; Oh, S. W.; and Kim, S. J. 2022. Error compensation framework for flow-guided video inpainting. In *European Conference on Computer Vision*, 375–390.

Ke, L.; Tai, Y.-W.; and Tang, C.-K. 2021. Occlusion-Aware Video Object Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14468–14478.

Kim, D.; Woo, S.; Lee, J.-Y.; and Kweon, I. S. 2019. Deep Blind Video Decaptioning by Temporal Aggregation and Recurrence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4263–4272.

Kim, D.; Woo, S.; Lee, J.-Y.; and Kweon, I. S. 2020. Recurrent Temporal Aggregation Framework for Deep Video Inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(5): 1038–1052.

Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M.-H. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 179–195.

Lee, S.; Oh, S. W.; Won, D.; and Kim, S. J. 2019. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4413–4421.

Li, A.; Zhao, S.; Ma, X.; Gong, M.; Qi, J.; Zhang, R.; Tao, D.; and Kotagiri, R. 2020. Short-Term and Long-Term Context Aggregation Network for Video Inpainting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 728–743.

Li, Z.; Lu, C.-Z.; Qin, J.; Guo, C.-L.; and Cheng, M.-M. 2022. Towards an End-to-End Framework for Flow-Guided Video Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17562–17571.

Liao, M.; Lu, F.; Zhou, D.; Zhang, S.; Li, W.; and Yang, R. 2020. Dvi: Depth guided video inpainting for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–17.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, 773–782.

Liu, R.; Deng, H.; Huang, Y.; Shi, X.; Lu, L.; Sun, W.; Wang, X.; Dai, J.; and Li, H. 2021. FuseFormer: Fusing Fine-Grained Information in Transformers for Video Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14040–14049.

Liu, R.; Li, B.; and Zhu, Y. 2022. Temporal Group Fusion Network for Deep Video Inpainting. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(6): 3539–3551.

Liu, R.; Weng, Z.; Zhu, Y.; and Li, B. 2020. Temporal Adaptive Alignment Network for Deep Video Inpainting. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 927–933.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 11(4): 674–693.

Pang, T.; Zheng, H.; Quan, Y.; and Ji, H. 2021. Recorrupted-to-Recorrupted: Unsupervised Deep Learning for Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2043–2052.

Perazzi, F.; Pont-Tuset, J.; Mcwilliams, B.; Gool, L. V.; and Sorkine-Hornung, A. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 724–732.

Ren, J.; Zheng, Q.; Zhao, Y.; Xu, X.; and Li, C. 2022. DLFormer: Discrete Latent Transformer for Video Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3511–3520.

Seoung, O., Wug; Sungho, L.; Joon-Young, L.; and Seon, K., Joo. 2019. Onion-Peel Networks for Deep Video Completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4402–4411.

Shang, Y.; Xu, D.; Zong, Z.; Nie, L.; and Yan, Y. 2022. Network binarization via contrastive learning. In *Proceedings of the European conference on computer vision (ECCV)*, 586–602.

Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; and Yan, Y. 2023. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1972–1981.

Shukla, T.; Maheshwari, P.; Singh, R.; Shukla, A.; Kulkarni, K.; and Turaga, P. 2023. Scene Graph Driven Text-Prompt Generation for Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 759–768.

Somani, A.; Banerjee, P.; Rastogi, M.; Agarwal, K.; Prasad, D. K.; and Habib, A. 2023. Image Inpainting With Hypergraphs for Resolution Improvement in Scanning Acoustic Microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3112–3121.

Srinivasan, V. S. R.; Ma, R.; Tang, Q.; Yi, Z.; and Xu, Z. 2021. Spatial-Temporal Residual Aggregation for High Resolution Video Inpainting. *arXiv preprint arXiv:2111.03574*.

Wang, C.; Huang, H.; Han, X.; and Wang, J. 2019. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI Conference on Artificial Intellignce (AAAI)*, 5232–5239.

Wu, Z.; Sun, C.; Xuan, H.; and Yan, Y. 2023a. Deep Stereo Video Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5693–5702.

Wu, Z.; Sun, C.; Xuan, H.; Zhang, K.; and Yan, Y. 2023b. Divide-and-Conquer Completion Network for Video Inpainting. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33(6): 2753–2766.

Wu, Z.; Zhang, K.; Sun, C.; Xuan, H.; and Yan, Y. 2023c. Flow-Guided Deformable Alignment Network with Self-Supervision for Video Inpainting. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wu, Z.; Zhang, K.; Xuan, H.; Yang, J.; and Yan, Y. 2021. DAPC-Net: Deformable Alignment and Pyramid Context Completion Networks for Video Inpainting. *IEEE Signal Processing Letters (SPL)*, 28: 1145–1149.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 603–619.

Xu, R.; Li, X.; Zhou, B.; and Loy, C. C. 2019. Deep Flow-Guided Video Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3723–3732.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *Proceedings of the European conference on computer vision (ECCV)*, 328–345.

Yu, Y.; Fan, H.; and Zhang, L. 2023. Deficiency-Aware Masked Transformer for Video Inpainting. *arXiv preprint arXiv:2307.08629*.

Yu, Y.; Zhan, F.; Lu, S.; Pan, J.; Ma, F.; Xie, X.; and Miao, C. 2021. WaveFill: A Wavelet-based Generation Network for Image Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14094–14103.

Zeng, Y.; Fu, J.; and Chao, H. 2020. Learning Joint Spatial-Temporal Transformations for Video Inpainting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3723–3732.

Zeng, Y.; Fu, J.; Chao, H.; and Guo, B. 2019. Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1486–1494.

Zhang, K.; Fu, J.; and Liu, D. 2022. Flow-Guided Transformer for Video Inpainting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 74–90.

Zhang, K.; Peng, J.; Fu, J.; and Liu, D. 2023. Exploiting Optical Flow Guidance for Transformer-Based Video Inpainting. *arXiv preprint arXiv:2301.10048*.

Zhang, K.; Wu, S.; Wu, Z.; Yuan, X.; and Zhao, C. 2022. Fractional Optimization Model for Infrared and Visible Image Fusion. In *Proceedings of the British Machine Vision Conference (BMVC)*, 1–12.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.

Zhang, Y.; Wu, Z.; and Yan, Y. 2023. PFTA-Net: Progressive Feature Alignment and Temporal Attention Fusion Networks for Video Inpainting. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 191–195.

Zou, X.; Yang, L.; Liu, D.; and Lee, Y. J. 2021. Progressive Temporal Feature Alignment Network for Video Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16448–16457.