IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2023, accepted 27 January 2024, date of publication 1 February 2024, date of current version 14 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3361283

RESEARCH ARTICLE

Deep Transformer Based Video Inpainting Using Fast Fourier Tokenization

TAEWAN KIM[1], JINWOO KIM[2], HEESEOK OH[3], AND JIWOO KANG[4,5], (Member, IEEE)

[1]Data Science Major, Dongduk Women's University, Seoul 02748, South Korea
[2]Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
[3]Department of Applied AI, Hansung University, Seoul 02876, South Korea
[4]Division of Artificial Intelligence Engineering, Sookmyung Women's University, Seoul 04310, South Korea
[5]Artificial Intelligence Innovation Research Center, Sookmyung Women's University, Seoul 04310, South Korea

Corresponding authors: Heeseok Oh (ohhs@hansung.ac.kr) and Jiwoo Kang (jwkang@sookmyung.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government Ministry of Science and ICT of South Korea (Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms, 100%) under Grant RS-2023-00229451; and in part by the Sookmyung Women's University Research Grants under Grant 1-2303-2015.

ABSTRACT Bridging distant space-time interactions is important for high-quality video inpainting with large moving masks. Most existing technologies exploit patch similarities within the frames, or leaverage large-scale training data to fill the hole along spatial and temporal dimensions. Recent works introduce promissing Transformer architecture into deep video inpainting to escape from the dominanace of nearby interactions and achieve superior performance than their baselines. However, such methods still struggle to complete larger holes containing complicated scenes. To alleviate this issue, we first employ a fast Fourier convolutions, which cover the frame-wide receptive field, for token representation. Then, the token passes through the seperated spatio-temporal transformer to explicitly moel the long-range context relations and simultaneously complete the missing regions in all input frames. By formulating video inpainting as a directionless sequence-to-sequence prediction task, our model fills visually consistent content, even under conditions such as large missing areas or complex geometries. Furthermore, our spatio-temporal transformer iteratively fills the hole from the boundary enabling it to exploit rich contextual information. We validate the superiority of the proposed model by using standard stationary masks and more realistic moving object masks. Both qualitative and quantitative results show that our model compares favorably against the state-of-the-art algorithms.

INDEX TERMS Video inpainting, video completion, free-form inpainting, object removal, adversarial learning.

I. INTRODUCTION

Video inpainting is the process of intelligently filling missing regions within a video frame, while maintaining spatial and temporal coherence. This task holds great significance across various real-world applications, encompassing restoration (eliminating permanent defects like scratches and dust), video re-touching (removing unwanted objects and water-marks), and stabilization (mitigating fluctuated motion and de-flickering). Nevertheless, achieving high-quality video inpainting remains a formidable challenge due to the absence of effective long-range interactions within the space-time domains.

Early patch-based video inpainting methods aimed to address the task by replacing the masked region with the most similar patch found elsewhere in the video [1], [2], [3]. However, these methods were often time-consuming and demonstrated limited capacity in synthesizing non-repetitive and intricate regions. This limitation arises from the assumption that there exists a hint for the missing portions within the observable regions. More recently, learning-based approaches have significantly elevated the performance

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

of video inpainting. These methods employ techniques like 3D convolutions and recurrent networks to improve results [4], [5], [6]. By aggregating information from neighboring frames, they endeavor to complete the gaps in the video. Among these advances, attention-based modules have proven highly effective, facilitating the transfer of long-range relationships between visible and obscured regions in the video [7], [8]. Despite these notable strides, the primary challenge of video inpainting remains the need to establish connections and effectively integrate visible information into synthesized contexts, all while considering inter-frame and intra-frame relationships.

In recent times, Transformers have gained significant prominence, emerging as the de-facto standard architecture for language-related tasks [9], [10], [11]. Notably, they have begun to demonstrate comparable or even superior performance to Convolutional Neural Networks (CNNs) across a diverse array of vision benchmarks [12], [13]. In contrast to CNN models, Transformers boast a robust representation capability and are characterized by their freedom from inductive biases. A key feature of Transformers is their ability to facilitate long-term interactions through the incorporation of dense attention mechanisms. This capacity has been leveraged in preliminary research efforts to model structural relationships in the context of natural image synthesis. This, in turn, leads to the generation of natural outputs through the optimization of underlying data distributions [14], [15], [16].

Motivated by the growing trend of employing transformer architecture in computer vision tasks, we present a novel, high-fidelity pluralistic video inpainting approach. In particular, our method treats video inpainting as a directionless sequence-to-sequence prediction task, effectively capturing both short- and long-term interactions through multi-head self-attention mechanisms. However, as highlighted in recent literature [17], [18], [19], transformers excel at capturing long-range interactions among input tokens but are less adept at capturing fine-grained local dependencies. Conversely, convolutional layers are adept at capturing local details but necessitate deeper layers to grasp the broader contextual understanding. This duality underscores the distinct limitations of transformers and CNNs.

In this study, we present a novel approach that leverages the strengths of both transformer and convolutional architectures. Our main insight is to harness the global structural dependencies through transformer layers, while using convolutional layers to enhance local texture contexts based on these global structural insights. However, directly applying transformer models to visual generation tasks poses challenges. Unlike natural language processing (NLP), where each word is treated as a vector for token embeddings, determining suitable token representations for visual tasks is less clear. Prior studies have resorted to considering every pixel or non-overlapping patches (e.g. 16×16) as token representations. Yet, due to high memory demands associated with longer input sequences, these methods suffer from resolution-related issues [20]. To address this concern, we incorporate convolutional layers to efficiently learn the compositional nature of masked video frames. Nonetheless, we've observed that conventional convolutional architectures may lack a sufficiently large receptive field for efficient token representations [13], [14], [16]. To overcome this limitation, we introduce a novel token representation approach based on recently developed fast Fourier convolutions (FFC) [21], [22]. This technique has a profound impact, allowing for frame-wise receptive fields that cover entire frames even in the initial layers of the network.

To address the computational complexity associated with self-attention, which grows quadratically with the frame length and becomes intractable for video transformers, we propose a spatially and temporally separated transformer backbone. This architecture aims to efficiently process a large number of spatio-temporal tokens that arise in videos. Specifically, we decouple the transformers across space-time volumes, enabling the coherent search for tokens from all frames and the simultaneous completion of all input frames. This design enables the model to synthesize stationary background textures within intra-frames and subsequently refine temporal consistency across inter-frames. We empirically assess this approach across various scalable transformer designs. Furthermore, to effectively complete intricate details even within large hole samples, we introduce an iterative refinement strategy. This involves gradually eroding the hole while refining tokens. Our design iteratively deduces and aggregates hole boundaries within the encoded feature map. Consequently, our network can tap into richer contextual information for the missing regions at each iteration.

Moreover, transformers are often considered ''data-hungry'' models due to their inductive bias-free nature, necessitating sufficiently large datasets for effective training. However, video datasets are typically relatively small in comparison. To address this challenge, we propose a strategy to effectively train transformer models on smaller video datasets by leveraging pre-training with larger image datasets. Our training approach harnesses a collection of static images to pre-train the proposed network. An example result depicting a successful content generation in a challenging object removal scenario is illustrated in Figure 1. Through an extensive series of experiments, we showcase that our model surpasses existing state-of-the-art approaches by a substantial margin in terms of metrics such as PSNR, SSIM, and VFID. Furthermore, we substantiate the efficacy of our proposed techniques through comprehensive ablation studies. In summary, our contributions can be encapsulated as follows.

1) We introduce a novel video inpainting network leveraging the advancements of Fast Fourier Convolutions (FFCs). These FFCs not only facilitate the incorporation of context-rich token representations but also contribute to refining local texture details, thereby significantly enhancing the overall network performance.
2) We put forth an innovative interwoven spatial-temporal transformer framework designed to proficiently

FIGURE 1. We introduce a video inpainting network based on transformers with an iterative refinement mechanism. Our model aims to complete the gray regions shown in the top row, and the synthesized frames are visualized in the bottom row (please zoom in for finer details).

capture global structural dependencies. This hierarchical transformer architecture empowers both intra- and inter-frame tokens to seamlessly engage with spatially and temporally coherent features, ultimately facilitating the restoration of the underlying global structure.

3) We present an iterative refinement module as part of our proposed methodology, aimed at further enhancing the accuracy of inpainting in deeper pixels of the holes. This module progressively accumulates richer contextual information for the regions with missing data at each step, contributing to the overall improvement of the inpainting results.

The structure of this article is as follows. Section II provides an overview of related work, encompassing image inpainting and video inpainting, in order to review the latest algorithms in these domains. In Section III, we delve into the comprehensive framework, outlining the background of Fast Fourier Convolutions (FFCs), the significance of Transformers, the architecture of our proposed model, and the intricacies of the training process. Moving forward, Section IV presents the database employed for performance evaluation and details the experimental results obtained. Finally, the paper concludes with a comprehensive summary in Section V.

II. RELATED WORKS

A. IMAGE INPAINTING METHODS

Traditional methods for image inpainting can be broadly categorized into diffusion-based [23], [24], [25] and patch-based approaches [26], [27], [28]. The former involves propagating texture from known regions to unknown (missing) regions and works well for small holes but tends to generate artifacts and noisy outcomes for larger holes. The latter, on the other hand, focuses on matching and copying nearest neighbor background patches. More recently, many researchers have turned to leveraging large image datasets for generating semantically coherent content using learning-based methods. Adversarial training, in particular, has been employed to enhance the realism of inpainted images [29], [30], [31], [32]. The concept of the context encoder was an early endeavor in generating reasonable results through feature learning [33]. Subsequent methods have aimed to enhance the visual quality of inpainted images to handle free-form masks, often adopting a two-stage refinement structure (such as coarse-to-fine architectures involving edges and structures) [34], [35], [36].

Building upon these foundations, numerous studies have explored the utilization of attention layers to learn correlations between background and foreground feature maps, allowing for the borrowing of pixels from distant locations [22], [37]. In pursuit of further refinement, image inpainting methods have embraced recursive hole-filling schemes to address larger holes. These methods ensure a confident region extending from the boundary to the center within feature spaces [38]. Our work builds upon attention and iterative refinement frameworks to tackle the challenges posed by the video inpainting task.

B. VIDEO INPAINTING METHODS

Video inpainting not only inherits the challenges faced by the image inpainting task but also introduces the need for time-consistent content generation. Early video inpainting methods often approached the space-time filling process through patch-based optimization techniques [26], [39], [40]. These methods completed holes by utilizing 3D patches in the spatio-temporal domain as synthesis units. For instance, Huang et al. proposed a non-parametric optimization approach that combined flow-field estimation and flow-guided patch synthesis [2], [3], [41]. While these methods yielded impressive outcomes, they often assumed stationary motion fields within holes and were constrained by dynamic camera motion, in addition to facing computational challenges.

In recent years, numerous researchers have turned to large datasets and deep learning models to generate plausible content. Early attempts such as combining 2D and 3D CNNs aimed to learn temporal and spatial features [6], although often resulting in blurry results. Inspired by flow-based methodologies [42], [43], [44], [45], Xu et al. explicitly estimated both appearance and optical flow to aid in propagating content from potentially distant frames. Kim et al. introduced a recurrent network to aggregate temporal features from nearby frames [5]. Chang et al. developed free-form video inpainting with 3D gated convolutions and temporal PatchGAN [4]. However, due to their limited ability to

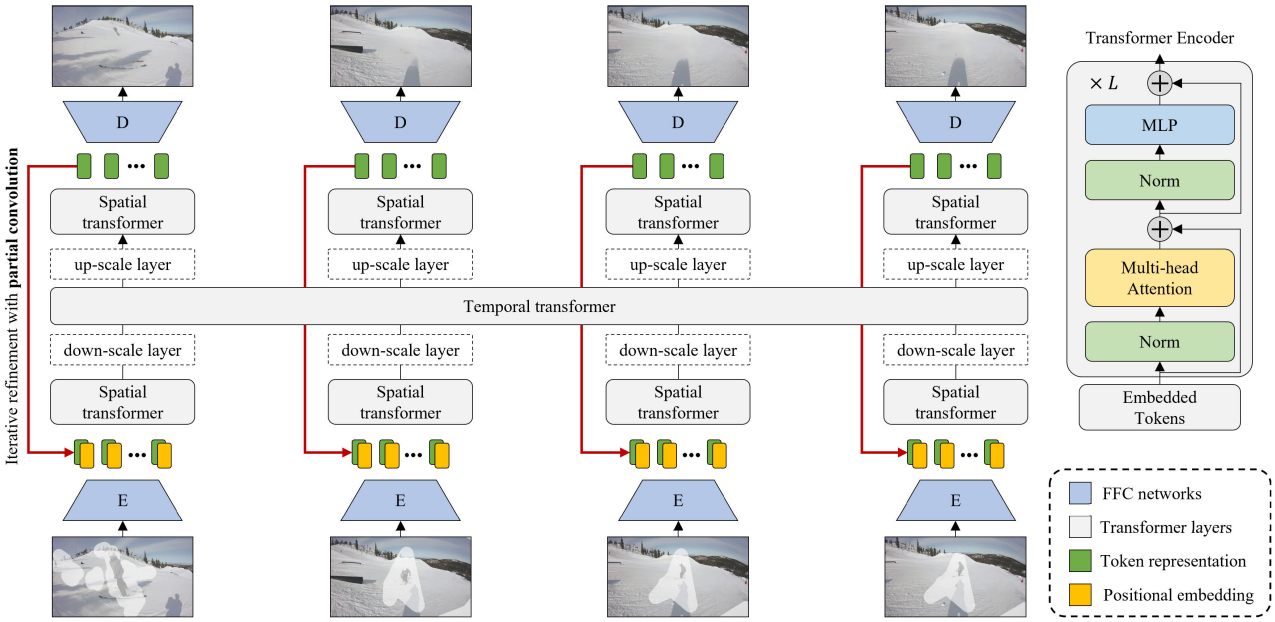

FIGURE 2. Overview of the Transformer-based Video Inpainting (TVI) architecture. TVI comprises separated spatial and temporal transformer blocks along with an iterative refinement module. The initial step involves embedding multiple input frames with independent tokens, followed by the concurrent hole-filling process through the transformer module to achieve globally coherent synthesis.

model long-range correspondences, these approaches often struggled to capture visible content from distant frames.

To address these limitations, recent approaches have embraced attention modules, showing promising performance. Oh et al. progressively filled missing regions from the boundary to the center with asymmetric attention to calculate similarities between target and reference frames [8]. Zeng et al. proposed STTN, directly applying multi-head self-attention to the video inpainting task, enabling the simultaneous completion of input frames while considering spatial-temporal similarity [46]. However, STTN introduced substantial computational demands. Specifically, applying a single multi-head attention layer to images with a resolution of 128×128 and 8 batches required more than 32GB of memory, which is generally impractical. Drawing inspiration from STTN, Liu et al. introduced DSTT, disentangling spatial and temporal learning tasks into two sub-tasks [47]. In contrast to the approach of DSTT, our method iteratively fills the missing hole in the feature domain, extracting continent tokens from each frame with smaller dimensions to propagate long-range interactions across space-time regions. Moreover, we take a step further in presenting an efficient training methodology for the data-hungry transformer architecture.

C. LIMITATIONS AND ADVANTAGES

Existing video inpainting methods exhibit varying limitations and advantages. The limitations often include sensitivity to certain types of scenes, computational inefficiency, and challenges in handling real-time processing. On the other hand, advantages may include superior performance in specific scenarios, efficient handling of dynamic content, or innovative approaches to addressing inpainting challenges. Our proposed method, detailed in this paper, leverages

TABLE 1. Comparison of existing articles in video inpainting.

Methods	Limitations	Advantages
Pathch-based optimization [26], [39], [40]	Subject to bluring on the borders of coherent inpainted regions	Capable of using arbitrary patch sizes
flow-guided patch synthesis [2], [3], [41]	Combinatorial labelling optimization computationally expensive. Lack of real-time processing	Presumably good for synthesising missing parts of structures
Gated Conv. [4]	Facing challenges in dealing with large holes	Robustly filling pixels for arbitrary masks
STTN [46] and DSTTN [47]	Facing challenges in handling complex scenes and dynamic content	Effective in capturing long-range dependencies
Our work	-	Superior performance in capturing distant space-time dependencies

Fourier Frame Convolution, Spatial and Temporal Transformers, and an Iterative Refinement process.

In summary, the field of video inpainting continues to evolve, with researchers exploring innovative techniques and applications. The comparison in Table 1 provides a snapshot of the diverse landscape of video inpainting methods and their respective strengths and weaknesses.

III. METHODS

A. OVERVIEW

1) PROBLEM FORMULATION

Let $X_1^T = \{X_1, X_2, \ldots, X_T\}$ represent a set of corrupted video frames with a sequence length T, and $M_1^T = \{M_1, M_2, \ldots, M_T\}$ be the corresponding frame-wise masks. Our objective is to learn a mapping function G that generates reasonable video output. This can be expressed as $G : X_1^T \rightarrow \hat{Y}_1^T$, where $\hat{Y}_1^T = \{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_T\}$ represents the predicted video frames. These predictions aim to closely approximate the target video frames $Y_1^T = \{Y_1, Y_2, \ldots, Y_T\}$.

To achieve this, we cast video inpainting as a multi-input and multi-output generative task. In this task, we aim to estimate the conditional distribution $p(Y_1^T | X_1^T)$. Our

motivation stems from the observation that a missing region in a current frame could potentially be revealed in both nearby and distant frames. For instance, if a mask is moving quickly, the missing region might be recoverable from adjacent frames by borrowing texture information. Conversely, when a mask is large and moving slowly, the occluded region could be visible in a distant frame.

In our approach, we leverage both adjacent and distant frames as conditions to simultaneously fill in the missing input frames. We adopt the Markov assumption [5], [46], which allows us to factorize the multiple conditional inputs and corresponding multiple outputs as a product form. Mathematically, this can be represented as:

$$p(\hat{Y}_1^T|X_1^T) = \prod_{t=1}^{T} p\left(\hat{Y}_t^{t+T_R}|X_t^{t+T_R}, X_{1,s}^T\right),$$

Here, $X_t^{t+T_R}$ represents the adjacent frames, and $X_{1,s}^T$ represents the distant frames. The distant frames are uniformly sampled from the total frames X_1^T with a sampling rate of s, following the approach in prior studies [46].

2) NETWORK DESIGN

In order to achieve spatially and temporally consistent content inpainting, we introduce a novel inference model termed **TVI** that utilizes a **T**ransformer-based **V**ideo **I**npainting architecture. An overview of our video inpainting framework is depicted in Figure 2. The TVI architecture is composed of a frame-level encoder-decoder and a cascaded spatial-temporal transformer.

The frame-level encoder is designed to capture the low-level structural information of the video frames. To achieve this, we leverage the **Fast Fourier Convolutions (FFCs)**, which allow efficient modeling of token representations while considering the entire frame-wide receptive field. Similarly, the frame-level decoder is responsible for transforming the encoded features back into the spatial domain, effectively completing the missing content.

The pivotal component of our architecture is the cascaded spatial-temporal transformer. This module is crucial for capturing long-range interactions across frames, enabling the restoration of global contextual information. To reinforce the spatial-temporal constraints and improve the overall quality of predictions, we introduce a plug-and-play recurrent feature reasoning process. This process facilitates the prediction of each global structure by iteratively inferring and accumulating hole boundaries within the encoded feature map.

By progressively strengthening the constraints governing spatial-temporal content, our proposed TVI model generates semantically coherent results that reflect the underlying context of the input video frames.

B. FAST FOURIER CONVOLUTION

1) BACKGROUND

Traditional fully convolutional models often struggle to achieve a large receptive field using small convolutional kernels (e.g., 3×3), leading to a need for deep networks with substantial memory demands. This limitation becomes particularly evident in video inpainting tasks involving large moving masks, where generators with insufficient receptive fields tend to observe only the neighboring missing pixels. Consequently, maintaining visually coherent contexts becomes challenging.

To address these limitations, we adopt the Fast Fourier Convolutions (FFCs) approach introduced in [21]. FFC provides an effective solution by leveraging local and non-local receptive fields in a single unit. The core idea of FFC is rooted in the channel-wise Fast Fourier Transform (FFT) [48], which significantly enlarges the image-wide receptive field, ensuring the consideration of global context for all layers.

FFC comprises two interconnected branches: a spatial (or local) branch that employs conventional convolutions on a portion of input feature channels, and a spectral (or global) branch that operates in the spectral domain using Real FFT to capture global context. These branches work in parallel, each with a different receptive field, allowing simultaneous acquisition of local and global information. The aggregation of features between these branches is performed internally within the FFC unit.

Formally, given an input feature volume $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote spatial resolution, the two branches are obtained by splitting the dimension along the channel axis, resulting in local and global features denoted as $\mathbf{X} = \{\mathbf{X}_l, \mathbf{X}_g\}$. The local feature $\mathbf{X}_l$, with dimensions $\mathbb{R}^{H\times W\times(1-\alpha_{in})C}$, focuses on learning local details using conventional convolution operations. On the other hand, the global feature $\mathbf{X}_g$, with dimensions $\mathbb{R}^{H\times W\times\alpha_{in}C}$, captures global context by transforming the spatial domain into the spectral domain using Real FFT. The parameter α_{in} controls the percentage of feature channels allocated to the global branch, varying between 0 and 1. The output features of the local and global branches are then aggregated to form the final feature volume $\mathbf{Y} = \{\mathbf{Y}_l, \mathbf{Y}_g\}$.

The FFC unit's internal operations can be expressed through equations:

$$\mathbf{Y}_l = \mathbf{Y}_{l\to l} + \mathbf{Y}_{g\to l} = f_l(\mathbf{X}_l) + f_{g\to l}(\mathbf{X}_g) \quad (1)$$

$$\mathbf{Y}_g = \mathbf{Y}_{g\to g} + \mathbf{Y}_{l\to g} = f_g(\mathbf{X}_g) + f_{l\to g}(\mathbf{X}_l), \quad (2)$$

where f_l, $f_{g\to l}$, and $f_{l\to g}$ represent convolution operations with 3×3 kernel shapes, with $f_{l\to g}$ reflecting the spectral transformation. This FFC architectural design, illustrated in Figure 3, closely follows the LaMa framework [22], employing a single Fourier unit.

2) FFC-BASED ENCODER

To harness the potent expressive capabilities of transformers for synthesis, we face the challenge of effectively representing each fixed-size frame ($432 \times 245 \times 3$) as an independent token. However, treating individual pixels as tokens for training the transformer becomes infeasible due to

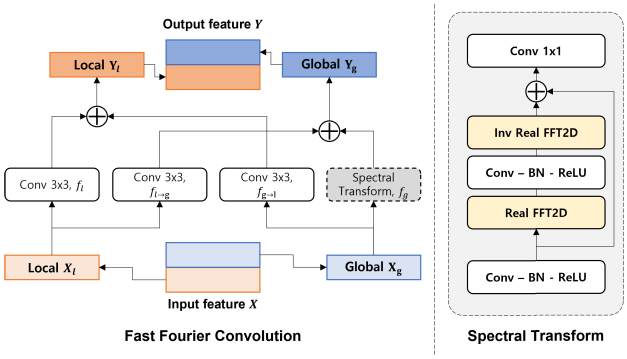

FIGURE 3. Illustration of Fast Fourier Convolution (FFC). The token representation is based on FFC which ensures a large receptive field and can avoid meaningless operation on the large hole regions. "⊕" denotes element-sise sum.

the substantial increase in sequence length (298, 080 tokens for each frame). In order to make it possible to input masked frames X_1^T into the transformer with manageable sequence lengths, we draw inspiration from ideas in neural discrete representation learning [14], [16] and integrate the representation abilities of Fast Fourier Convolutions (FFCs).

Specifically, we design the frame-level encoder by stacking multiple FFC layers with down-sampling. This architecture aims to capture both global context and local details from early layers, which is essential for creating a compact token representation. Following this, we obtain spatial features along the temporal index T_R, resulting in a size of $H \times W \times C \times T_R$. These spatial features are then flattened for each frame, yielding 1D sequences of size $(HW \times C) \times T_R$. In our implementation, we set H, W, C, and T_R to 30, 54, 512, and 4, respectively. The effectiveness of our approach is empirically demonstrated in both quantitative and qualitative evaluations in Section IV-D.

Note: Please replace "Section IV-D" with the correct reference to the relevant section in your article.

3) FFC-BASED DECODER

Following token representation, the acquired tokens are input into a novel transformer architecture designed for comprehensive information aggregation. These processed tokens are subsequently translated to target frames via a frame-level decoder. To ensure the synthesis of photorealistic content through gradual up-sampling, we opt for Fast Fourier Convolution (FFC) layers as our frame-level decoder, which employs groups of dilated convolutions [29]. Notably, due to their ability to cover the entire image with an expansive receptive field, FFCs outperform recent CNN-based architectures in terms of performance.

C. CASCADED SPATIAL AND TEMPORAL TRANSFORMER

1) BACKGROUND

We adopt the transformer encoder architecture proposed by Dosovitskiy et al. [13] as our fundamental building block. Here, we provide a brief overview of the transformer's functionality. The primary operation carried out in this layer is self-attention, which is computed over a sequence of tokens. As illustrated in Figure 2 (right), the transformer encoder comprises alternating layers of multi-head self-attention (MSA), responsible for capturing long-range dependencies, and multi-layer perceptron (MLP) blocks with GELU non-linearity. Layer normalization (LN) is applied before both components, and each block employs a residual connection. These operations are represented as follows:

$$\mathbf{z}_0 = \left[\mathbf{x}^1; \mathbf{x}^2; \ldots; \mathbf{x}^M\right] + \mathbf{E}_{pos}, \tag{3}$$

$$\mathbf{z}'_l = \mathbf{MSA}\left(\mathbf{LN}\left(\mathbf{z}_{l-1}\right)\right) + \mathbf{z}_{l-1}, \tag{4}$$

$$\mathbf{z}_l = \mathbf{MLP}\left(\mathbf{LN}\left(\mathbf{z}'_l\right)\right) + \mathbf{z}'_l, \tag{5}$$

where $\mathbf{z} \in \mathbb{R}^{M\times C}$ represents the 1D sequence of M tokens $\mathbf{x}$ with C dimensions, and $\mathbf{E}_{pos} \in \mathbb{R}^{M\times C}$ denotes the position embeddings.

2) TRANSFORMER FOR SPATIO-TEMPORAL INTERACTION

We introduce a transformer-based architecture to capture coherent contents by processing all the represented tokens. As depicted in Figure 2, our model comprises three distinct transformer encoders in sequence. Similar to the design of BERT [49], the spatial transformer encoder takes token embeddings as inputs and computes the relationships between each token of the same temporal index. The representation for each temporal index is denoted as $\mathbf{x}_s^m \in \mathbb{R}^{HW\times C}$, where $m = 1, 2, \ldots, T_R$.

To account for temporal relationships, the tokens from the spatial transformer encoder are reshaped along the temporal dimension, $HW \times C \times T_R \rightarrow (HWT_R) \times C$. However, this leads to a significant increase in sequence length with the growing number of input frames T_R, resulting in computational complexity. To address this issue, we incorporate down-scaling and up-scaling layers before and after the temporal transformer encoder. Specifically, the down-scaling layer reshapes the 1D sequence of token embeddings into a 2D feature map $\mathbf{x}_s^m \in \mathbb{R}^{H\times W\times C}$ and applies stacked convolutions with a down-sampling module $\mathbf{x}_s^m \downarrow \in \mathbb{R}^{H/2\times W/2\times C}$. Subsequently, the 2D feature map is reshaped back into a 1D sequence of embedding tokens, yielding $\mathbf{x}_t \in \mathbb{R}^{(\frac{H}{2}\frac{W}{2}T_R)\times C}$. The temporal transformer encoder then processes the temporally grouped tokens $\mathbf{x}_t$ and computes relationships between each token recursively. Similarly, the up-scaling layer is designed to reshape the temporally computed tokens back to the dimensions of $(HW \times C) \times T_R$. Finally, the spatial transformer encoder is employed once more to further enhance the quality of synthesis.

3) INFERENCE VIA ITERATIVE REFINEMENT

To further enhance the quality of our model's output, we introduce an iterative refinement process within the transformer block, gradually improving the internal content. Differing from existing iterative techniques [8], our proposed model performs this refinement within the encoded feature space. This approach not only makes efficient use of

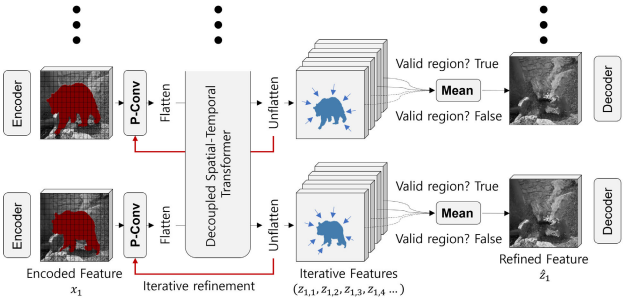

FIGURE 4. Illustration of the iterative refinement procedure. The area identification process is performed by partial convolution and the hole region gradually decreases during several times reasoning (blue arrows). After iterative refinements, collected feature maps are adaptively merged considering the valid region of the mask.

parameters, resulting in a lighter model, but also ensures superior performance.

In each iteration, a partial convolution [50] is employed as a fundamental module to determine the area that needs updating. This operation updates the mask and renormalizes the feature map post convolution calculation. Let $\mathbf{W}$ represent the convolutional kernel and $\mathbf{b}$ be the corresponding bias. The feature map $\mathbf{x}^*$ obtained from the partial convolution layer is given by:

$$\mathbf{x}^* = \begin{cases} \mathbf{W}^T\left(\mathbf{x} \odot \mathbf{m}\right)\dfrac{\text{if sum}(\mathbf{1})}{\text{sum}(\mathbf{m})} + \mathbf{b}, & \text{sum}(\mathbf{m}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\mathbf{x}$ and $\mathbf{m}$ represent the feature values for the current convolution window and the corresponding binary mask, respectively. Similarly, the updated mask value can be expressed as:

$$\mathbf{m}^* = \begin{cases} 1, & \text{if sum}(\mathbf{m}) > 0 \\ 0. & \text{otherwise} \end{cases} \tag{7}$$

By employing these equations, new masks are generated in which the holes become progressively smaller with each partial convolution layer.

Following multiple refinement iterations within the transformer, intermediate features are merged to prevent gradient vanishing issues, as discussed in prior works [38]. Instead of directly passing the last features to the decoder, we employ an adaptive merging approach that normalizes the value across the newly completed regions [38]. Let $\mathbf{z}_{m,n}$ denote the features from the n^{th} iteration, calculated along the temporal index $m = 1, 2, \ldots, T_R$. The value within the refined feature map $\hat{\mathbf{z}}_{m,n}$ is defined as:

$$\hat{\mathbf{z}}_m = \sum_{n=1}^{N} \frac{\mathbf{z}_{m,n}}{\mathbf{m}^*_{m,n}}, \tag{8}$$

where N represents the number of iterations. This approach enables the model to merge an arbitrary number of feature maps, ensuring the potential quality of synthesis. The details of the iterative refinement pipeline of our module are illustrated in Figure 4.

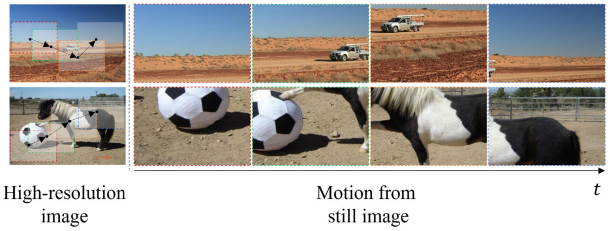

FIGURE 5. Examples of cropped images (right) from the high-resolution image (left). Cropped images imitate small to large motions.

D. TRAINING

1) LOSS FUNCTION

The loss function is formulated to address pixel-wise reconstruction accuracy, perceptual similarity, and temporal consistency. This involves minimizing the $L1$ distance between the generated and ground-truth frames to ensure pixel-wise reconstruction. The pixel loss components are defined as follows:

$$\mathcal{L}_{hole} = \left\| (1 - M_1^T) \odot \left(\hat{Y}_1^T - Y_1^T\right) \right\|, \tag{9}$$

$$\mathcal{L}_{valid} = \left\| M_1^T \odot \left(\hat{Y}_1^T - Y_1^T\right) \right\|. \tag{10}$$

Additionally, we incorporate the Structural Similarity Index Measure (SSIM), a perceptually motivated loss [51]:

$$\mathcal{L}_{SSIM} = \sum_{t=1}^{T} \text{SSIM}\left(\hat{Y}_t, Y_t\right). \tag{11}$$

To maintain temporal consistency, we adopt a Temporal Patch GAN as our discriminator [46]. We maintain the discriminator architecture and loss function as originally defined. This adversarial loss contributes to generating plausible and coherent results in video inpainting. The optimization function for the discriminator is as follows:

$$\begin{aligned} \mathcal{L}_D = {} & E_{x \sim P_{Y_1^T}(x)}\left[\text{ReLU}\left(1 - D(x)\right)\right] \\ & + E_{x \sim P_{\hat{Y}_1^T}(x)}\left[\text{ReLU}\left(1 + D(x)\right)\right]. \end{aligned} \tag{12}$$

Subsequently, the corresponding adversarial loss for the Transformer-based Video Inpainting (TVI) is defined as follows:

$$\mathcal{L}_{adv} = -E_{z \sim P_{\hat{Y}_1^T}}\left[D(z)\right]. \tag{13}$$

Finally, the comprehensive loss function is expressed as follows:

$$\mathcal{L} = \lambda_{hole} \cdot \mathcal{L}_{hole} + \lambda_{valid} \cdot \mathcal{L}_{valid} + \lambda_{SSIM} \cdot \mathcal{L}_{SSIM} + \lambda_{adv} \mathcal{L}_{adv}, \tag{14}$$

where hyperparameters are determined empirically (for instance, λ_{hole}, λ_{hole}, and λ_{hole} are all set to 1, and λ_{adv} is set to 0.1).

Algorithm 1 Training of Our Proposed Network

Inputs : $X_{1:T} : \{X_1, \ldots, X_T\}$, Corrupted frames;
$M_{1:T} : \{M_1, \ldots, M_T\}$, Frame-wise masks;
Outputs: $\hat{Y}_{1:T} : \{\hat{Y}_1, \ldots, \hat{Y}_T\}$, Outputs of the TVI;

initialization;
$\mathbf{x}_{1:T} \leftarrow$ FFC Encoder(X_1^T, M_1^T);
$\mathbf{z}_{1:T} \leftarrow$ PositionalEncoding$(\mathbf{x}_1^T)$;
$\mathbf{m}_{1:T} \leftarrow$ DownSampling$(M_{1:T})$;
$i \leftarrow 0$;
while *i smaller than N* **do**
  $\mathbf{z}_{1:T}^{i+1}, \mathbf{m}_{1:T}^{i+1} \leftarrow$ PartialConv$(\mathbf{z}_{1:T}^{i}, \mathbf{m}_{1:T}^{i})$;
  $\mathbf{z}_{1:T}^{i+1} \leftarrow$ SpatialTemporalTransformer$(\mathbf{z}_{1:T}^{i+1})$;
  FeatureGroup $\leftarrow$ FeatureGroup $+ \{\mathbf{z}_{1:T}^{i+1}\}$;
  $i \leftarrow i + 1$;
end
$\mathbf{x}_{1:T}^{merged} \leftarrow$ FeatureMerge(FeatureGroup);
$\hat{Y}_{1:T} \leftarrow$ FFCDecoder$(\mathbf{x}_{1:T}^{merged})$;
Updating the TVI with loss $\mathcal{L}$;

2) PRE-TRAINING FROM IMAGE DATASET

Transformer-based architectures are often considered "data-hungry," implying their effectiveness when abundant training data is available. Nonetheless, video datasets are relatively small compared to images, making our model susceptible to biases due to the limited training samples. To mitigate this concern, our training approach incorporates a collection of static images to pre-train the Transformer-based Video Inpainting (TVI) model.

Specifically, we utilize extensive high-resolution image datasets like Places2 [52] for training. In this process, we crop a single image while considering motion components. Drawing inspiration from optical flow methods [53], which assume smooth motion transitions in real-world videos, we simulate motion from a static image as follows:

$$\begin{aligned} c_x^{m+1} &= c_x^m + w \cdot \Delta x, \\ c_y^{m+1} &= c_y^m + h \cdot \Delta y, \end{aligned} \tag{15}$$

Here, (c_x^m, c_y^m) denotes the center point of cropped images, while w and h represent width and height, respectively. The variables Δx and Δy are random samples from a zero-centered normal distribution, serving to capture positional changes. Figure 5 illustrates examples of the cropped images used for our pre-training strategy.

E. IMPLEMENTATION DETAILS

Our FFC-based encoder and decoder layers draw inspiration from the ResNet architecture [54], where the CNN layers in the residual block are substituted with FFC layers. In our TVI model, we employ 3 downsampling blocks and 3 upsampling blocks. Our transformer model follows the ViT architecture, with the capacity being primarily adjusted by varying the number of stacked layers. We quantitatively discuss the transformer's capacity in Section IV-D. As our discriminator, we select the Temporal PatchGAN (T-PatchGAN) [4], comprising six layers of 3D convolutional layers. This module serves to classify whether each spatial and temporal feature is real or fake, similar to the standard GAN framework. This adversarial training mechanism encourages TVI to focus on spatial details and the temporal coherence of actual videos [4], [55]. Additionally, we manually set the recurrence number N to 8 for the transformer module to simplify training. The network's operational procedure is outlined in Algorithm 1.

Throughout the experiments, frames with a resolution of 432×240 are utilized for training the proposed model. The color values of all frames are linearly scaled to the range $[-1, 1]$. Prior to the training process, we initialize all network weights using the normalized distribution $\mathcal{N}(0, 1)$. Optimization is carried out using the Adam optimizer [56], with $(\beta_1, \beta_2) = (0.0, 0.99)$ applied to both TVI and the discriminator. We set the learning rate to a fixed value of $\lambda = 1e^{-4}$. To enhance model stability, we employ spectral normalization (SN) [57], which scales down weight matrices with their largest singular values. Our model is trained using a batch size of at least 2, leveraging a GPU with 128GB VRAM. However, we typically conduct training across more than 8 GPUs, accumulating VRAM up to 96GB. If hardware capabilities permit, 16-bit precision training is employed.

IV. EXPERIMENTS

In this section, we commence by introducing the datasets employed for validating the model, followed by an explanation of the training particulars for each dataset to ensure result reproducibility. We subsequently conduct a comprehensive evaluation of our approach, featuring both quantitative and qualitative analyses. These analyses encompass comparisons with recent video inpainting methods, alongside a user study. Furthermore, we carry out ablation experiments to assess the impact of various baseline components and present supplementary results.

A. DATASETS

Our comparative analysis encompasses two widely employed datasets extensively used in video inpainting research: Youtube-VOS [59] and DAVIS [60]. The Youtube-VOS dataset comprises 4, 453 videos spanning diverse scenes. The dataset is partitioned into train/validation/test subsets, with a distribution of 3, 471, 474, and 508 videos respectively. For the Youtube-VOS dataset, we adhere to the original dataset split and present experimental findings on the designated test set. The average video length within the Youtube-VOS dataset is approximately 150 frames. On the other hand, the DAVIS dataset encompasses 150 high-quality videos featuring dynamic camera and foreground motions. In accordance with established evaluation practices [43], we use 60 sequences for training and 90 sequences for testing.

Furthermore, to address the data-hungry nature of our model, we incorporate the high-resolution Places2 dataset [52] for pre-training. This dataset, tailored for

TABLE 2. Quantitative comparisons on two datasets using object mask, curve mask, and stationary mask. The best measures are in bold. † Lower value is better. ∗ Higher value is better.

	Youtube-Vos									DAVIS								
	Object mask			Curve mask			Stationary mask			Object mask			Curve mask			Stationary mask		
	PSNR*	SSIM*	VFID†	PSNR*	SSIM*	VFID†	PSNR*	SSIM*	VFID†	PSNR*	SSIM*	VFID†	PSNR*	SSIM*	VFID†	PSNR*	SSIM*	VFID†
OPN	33.53	0.8844	0.7618	34.16	0.9125	0.6602	36.15	0.9540	0.4004	32.91	0.8635	0.3664	33.78	0.9105	0.2701	36.33	0.9596	0.1281
CPN	33.18	0.8764	0.8257	32.88	0.8676	0.8841	35.86	0.9485	0.4606	32.60	0.8452	0.4331	32.47	0.8496	0.4802	36.55	0.9547	0.1637
FGVC	33.13	0.8832	0.7640	34.14	0.9212	0.640	35.09	0.9422	0.4017	31.95	0.8323	0.4010	32.84	0.8841	0.3432	33.92	0.9212	0.1734
STTN	34.86	0.9047	0.7276	36.07	0.9411	0.6136	39.60	0.9716	0.3132	33.60	0.8708	0.3831	34.83	0.9251	0.2882	38.78	0.9690	**0.1197**
FGT	**35.81**	0.9140	0.6113	37.63	**0.9712**	0.3668	**41.80**	0.9728	0.2923	33.29	0.8803	0.3492	**37.84**	**0.9630**	**0.2048**	41.95	0.9632	0.1474
TVI	35.57	**0.9166**	**0.6072**	**37.88**	**0.9631**	**0.3650**	39.87	**0.9741**	**0.2869**	**34.84**	**0.8832**	**0.3403**	35.93	0.9371	0.2207	**42.21**	**0.9739**	0.1208

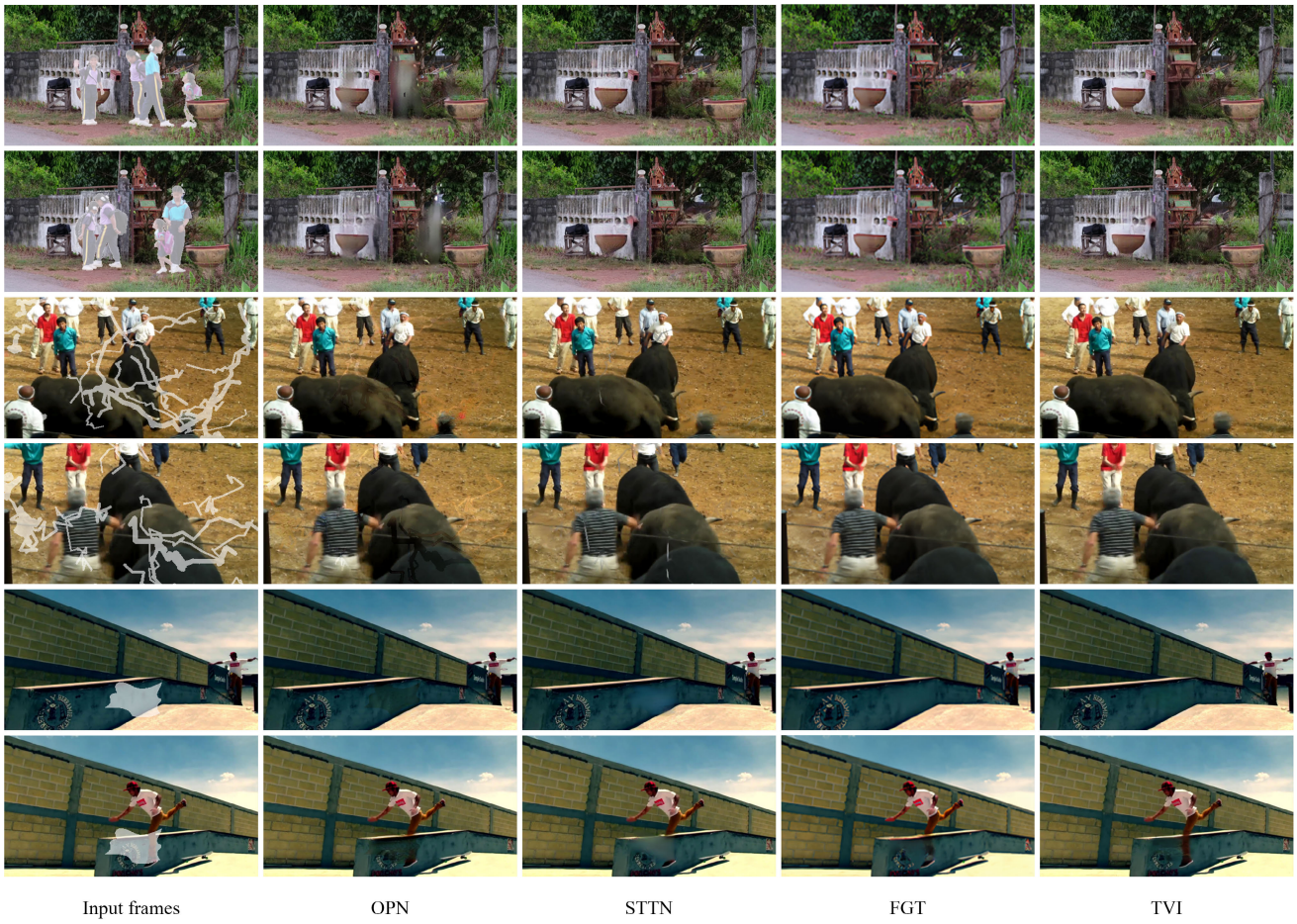

FIGURE 6. Qualitative comparisons of our methods with OPN [8], STTN [46], and FGT [58]. Our model generates globally coherent content than other benchmarks.

natural synthesis tasks, serves to imbue the model with prior knowledge through an inductive bias-free design. Additionally, we simulate real-world application scenarios by utilizing previously introduced image corruption methods [61]. Specifically, we apply three types of free-form masks: moving object-like masks, moving curve masks, and stationary masks.

It's noteworthy that our training strategies are slightly adapted based on the specific datasets. Given the limited size of the training video datasets, we initiate the training process using the high-resolution Places2 dataset. In this phase, we solely train the generator with the appearance loss term for 300 epochs. Subsequently, we incorporate the discriminator with adversarial loss to fine-tune the TVI model on both the Youtube-VOS and DAVIS datasets. The fine-tuning stage encompasses an additional 200 epochs.

B. BASELINES AND EVALUATION METRICS

Our evaluation includes a comparison between our proposed model and four existing deep-learning based video inpainting methods, namely OPN [8], CPN [7], FGVC [43], STTN [46], and FGT [58]. To ensure a fair assessment, we retrain these baseline models until convergence following the experimental settings detailed in each respective study. Below are the details of the compared baselines:

TABLE 3. User study and model efficiency on different methods. B is short of Billion.

	TVI WinRate (User study)			Model efficiency		
	Object mask	Curve mask	Stationary mask	Params	FLOPs	FPS
OPN	81.30%	64.32%	73.58%	12**M**	367B	12.7
STTN	69.48%	56.58%	66.28%	26M	233B	24.3
FGT	53.87%	58.41%	62.32%	38M	186B	36
TVI	-	-	-	18M	**96B**	**37.9**

- OPN: This model employs iterative refinement and incorporates attention modules in intermediate layers.
- CPN: CPN can compute affine matrices by fusing reference frame features based on image similarity.
- FGVC: FGVC addresses the limitations of existing flow-based video completion algorithms by utilizing flow-edge, non-local flow, and seamless blending modules.
- STTN: STTN learns joint spatial and temporal attention modules through multi-scale patch-based video frame representations.
- FGT: FGT leverage the motion discrepancy exposed by optical flows to instruct the attention retrieval in transformer for high fidelity video inpainting.

Quantitative comparisons are conducted using peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Video Fr'echet Inception Distance (VFID) [62]. While PSNR and SSIM assume pixel-wise independence, potentially favoring perceptually suboptimal results, VFID offers a more reliable perceptual evaluation. VFID calculates the distance between features using a pre-trained I3D model [63]. It's important to note that VFID scores are typically lower for our approach due to their reliance on the completed video, which primarily comprises original parts.

C. PERFORMANCE EVALUATION

1) QUANTITATIVE COMPARISON

Our quantitative evaluation involves various masks, including object, curve, and stationary masks, on both the Youtube-VOS and DAVIS datasets. The results, presented in Table 2, demonstrate the superior video completion performance of our model in comparison to state-of-the-art algorithms [7], [8], [43], [46], [61], particularly when considering the object mask for all evaluation metrics. Additionally, our approach exhibits potentially superior or competitive performance across other mask types in all evaluations. These findings underscore the critical role of our proposed module in enhancing the visual quality of inpainted videos.

Furthermore, we provide a comprehensive analysis of running time efficiency, as outlined in Table 3. Remarkably, our model showcases the lowest FLOPs and the highest FPS, signifying its exceptional efficiency in the context of video inpainting tasks.

2) QUALITATIVE COMPARISON

To underscore the superiority of our proposed method, we present selected results that highlight the capacity of our approach to address both short-range and long-range interactions in video inpainting scenarios. Figure 6 showcases video inpainting samples involving object removal, curve masks, and stationary mask corruptions. In all of these instances, our inpainting results exhibit remarkable coherence in both spatial and temporal aspects across various mask types. Notably, our model excels in synthesizing sharp and clear appearances during object removal tasks, effectively preserving background textures in regions that were previously invisible or occluded. We also provide additional results for further illustration, available in Section IV-D.

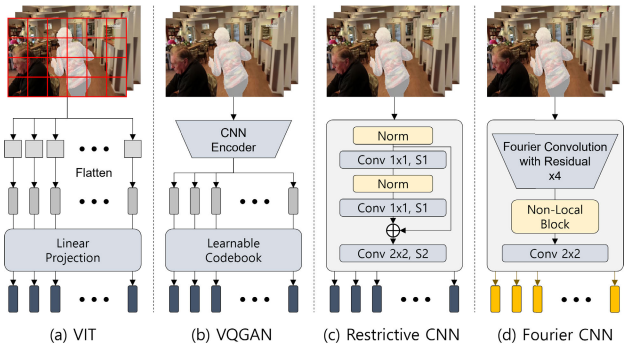

FIGURE 7. Visualization of token representation. (a) Patch-based token representation [13]. (b) Discrete feature to token [14]. (c) Restricted receptive field feature to token [16]. (d) Fast Fourier convolution-based token representation.

TABLE 4. Comparisons with different configurations of video inpainting architectures for the object removal task.

Method	Youtube-Vos			DAVIS		
	PSNR	SSIM	VFID	PSNR	SSIM	VFID
A Traditional Convolution	30.08	0.7718	0.8921	25.83	0.7859	0.5836
B + VIT	32.52	0.7732	0.8639	26.02	0.7937	0.5517
C + VQGAN	29.17	0.7950	0.8595	26.69	0.8268	0.5562
D + Restrictive CNN	31.29	0.8352	0.8525	27.58	0.8314	0.5419
E + FourierCNN	33.58	0.8421	0.8184	29.81	0.8381	0.4602
F + Spatial Transformer	33.62	0.8780	0.7709	31.80	0.8427	0.4153
G + Temporal Transformer	34.81	0.8917	0.7490	32.47	0.8649	0.3980
H + Decoupled Transformer	34.96	0.9158	0.6138	33.81	0.8733	0.3764
I + Pre-training from image dataset	**35.57**	**0.9166**	**0.6072**	**34.84**	**0.8832**	**0.3403**

3) USER STUDY

To mitigate potential biases inherent in selected evaluation metrics, we conducted a user study to assess the visual quality of our model in comparison to strong baseline methods such as OPN, STTN, and FGT. For the study, we randomly selected 20 videos from the DAVIS test split and introduced object, curve, and stationary masks to these samples. Subsequently, we used the baseline models to complete the corrupted videos. Paired comparisons were then carried out between our method (TVI) and the selected baselines, using the same set of videos.

The user study engaged 23 participants, each tasked with choosing the more plausible or visually natural video between our results and a randomly chosen counterpart from the baselines. Our method achieved the majority of votes over all baselines. Specifically, the win-rates of TVI against the baselines are as follows: OPN (81.3%), STTN (69.48%), and FGT (53.87%) for the object mask. This outcome underscores

FIGURE 8. Qualitative comparisons of different token representations.

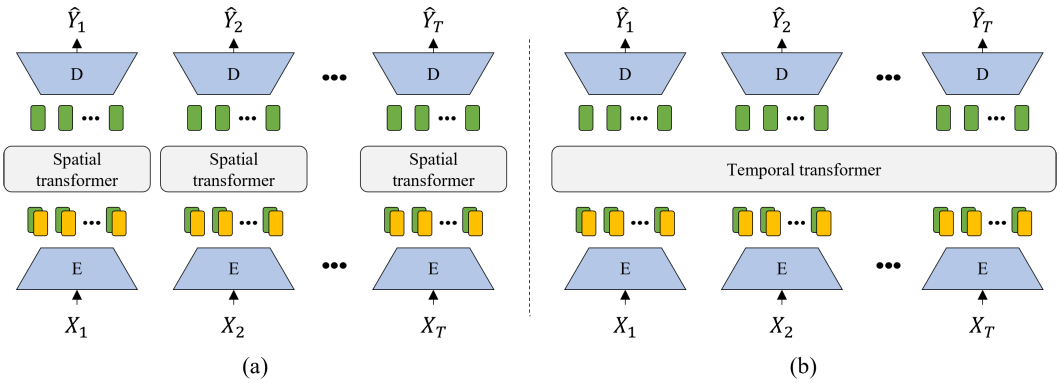

FIGURE 9. Architecture selections. (a) consists of a spatial transformer that interacts with frame-wise information. (b) consists of a temporal transformer that interacts with sequential frame information simultaneously.

that the video inpainting produced by our method is generally preferred and less discernible compared to the other baseline approaches.

D. ABLATION STUDIES

In this section, we conduct a series of ablation studies to individually assess the impact of various components in our proposed approach. We aim to analyze the effectiveness of Fourier convolutions, the role of transformer blocks, the benefits of iterative refinement, and the effects of pre-training from an image dataset.

1) THE POWER OF FOURIER CONVOLUTION

Fourier convolutions, rooted in periodic convolutions, offer full differentiability and the flexibility to seamlessly integrate and interchange with conventional convolutions. Owing to their comprehensive receptive field that spans the entire frame in the spectral domain, Fourier convolutions encourage the network to capture global context from the inception of the layer. This characteristic holds significance for our video inpainting framework, wherein frames are represented as sequential tokens due to their sensitivity to large and moving masks.

To underscore the potency of Fourier convolutions, we undertook experiments to analyze different token representation methods, drawing from recent vision transformer research [13], [14], [16]. As depicted in Figure 7, we divided each frame into fixed patches and flattened each patch to serve as a token, as per the approach of Vision Transformers (ViT) [13]. However, as indicated in Table 4 and Figure 8,

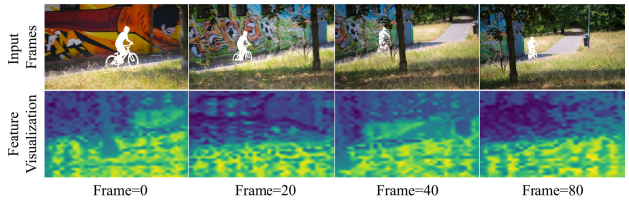

FIGURE 10. The visualization of input images and intermediate features illustrates consistent representation for distant frames with similar contextual meanings.

the results yielded temporally coherent appearances yet exhibited quantitatively and qualitatively blurry textures. Employing token representation akin to VQGAN [14], which initially encodes the image using conventional convolution layers and then quantizes its feature into tokens through a learnable dictionary, yielded visually plausible outcomes but struggled to reconstruct fine details. To extend the token representation comparison, we utilized the constrained CNN approach [16], designed to ensure each token encapsulated individual information without becoming intertwined with neighboring pixels. This method delivered relatively better quantitative and qualitative outcomes, though certain details remained deficient.

We attribute this discrepancy to the limited receptive field of token representation. Unlike other low-level vision tasks (e.g., style transfer, color transfer, super-resolution, etc.), in the context of video inpainting, a substantial portion of each frame is marred by the mask, leading to a dearth of information in those regions. Consequently, a restricted receptive field for autonomous token representation can result

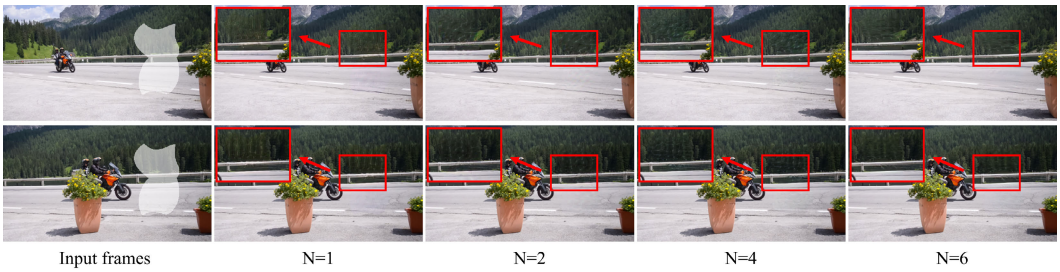

FIGURE 11. Comparison results for different iteration numbers.

in tokens that hold little utility within the video inpainting domain. Moreover, while transformer approaches excel at modeling non-local interactions, they tend to be less adept at capturing intricate local details. This accentuates the significance of fine-grained token representation in video inpainting tasks.

In contrast to preceding token representation strategies [13], [14], [16], our Fourier convolution-based token representation encompasses the global context, effectively bridging the gap between global and fine-grained token representation. As delineated in Table 4 and Figure 8, our outcomes showcase substantial enhancements in both quantitative and qualitative assessments.

2) EFFECTIVENESS OF THE SEPARATED SPATIAL AND TEMPORAL TRANSFORMER

To determine the optimal architecture configuration, we conducted a thorough validation of two plausible baseline models, as depicted in Figure 9 (a) and (b). These two baseline networks differ in terms of their interaction range. In Figure 9 (a), the spatial transformer operates on a sequence of individual frame tokens with positional embeddings, focusing on intramodal relationships within a frame. Consequently, this network reconstructs frames without accounting for temporal dependencies. Conversely, in Figure 9 (b), the temporal transformer extends local interactions into global interactions. This is achieved by taking an entire sequence of frame tokens as input and calculating dot-product attention across the sequence. In contrast to these models, our approach segregates short- and long-range visual dependencies.

Quantitative performance results for three distinct network configurations—F, G, and H—are presented in Table 4. Notably, the configuration I outperforms the other architectures across all metrics by a significant margin. This stems from the heightened capacity to capture spatial and temporal coherence features, which is facilitated by the separate handling of interaction modes. It's important to note that configurations F and G were equipped with the same number of transformer layers to mitigate performance discrepancies based on layer depth. This suggests that the process of identifying spatial dependencies is not only memory-efficient by circumventing long-range interactions but also excels in preserving finer details.

Additionally, we present intermediate results after passing through the temporal transformer. To achieve this, we applied one-step PCA to the features at intermediate stages. As shown in the figure 10, despite a significant gap between frames, parts with contextually similar meanings exhibit similar score values. This observation indicates the effective filling of masked regions through long-range feature interaction.

3) EFFECTIVENESS OF THE ITERATIVE REFINEMENT

The effectiveness of the iterative refinement module, which progressively enhances the visible region at the feature levels, constitutes a significant contribution of our work. In this section, we focus on elucidating the impact of the iterative refinement process. The results corresponding to different iteration values (N) on the DAVIS dataset are presented in Figure 11. These quantitative scores were obtained after the same number of training iterations. Notably, this ablation study underscores the robustness of our approach to variations in this hyperparameter. The findings also demonstrate an enhancement in performance with an increasing number of iterations. However, once the iteration count exceeds 6, the magnitude of performance improvement diminishes considerably. The detailed results subsequent to the integration of iterative refinement are showcased in Figure 11.

4) EFFECTIVENESS OF PRE-TRAINING FROM IMAGE DATASET

The effectiveness of pre-training the proposed model from large image datasets is notable in enhancing the learning of short- and long-term dependencies, a critical aspect of our transformer-based architecture. The inductive-free design of the transformer blocks necessitates a substantial amount of data for effective learning. In order to address this challenge, we employ pre-training with large-scale image datasets. Configuration F in Table 4 showcases the quantitative results of this pre-training approach. The TVI model, augmented with the pre-training procedure, exhibits improved performance in terms of PSNR (34.84), SSIM (0.8832), and VFID (0.3403) on the DAVIS dataset.

V. CONCLUSION

In summary, this paper presents an innovative transformer-based approach for video inpainting, specifically addressing the challenge of handling distant space-time visual dependencies. Leveraging the expressive power of Fourier

Frame Convolution (FFC), our method extracts tokens from sequential frames. These tokens undergo processing through dedicated spatial and temporal transformers, facilitating interframe completion and subsequent intra-frame coherence refinement. While our architectural design demonstrates effectiveness in establishing meaningful connections between distant frames, it is essential to consider future directions and potential drawbacks. Future research could explore optimizing computational efficiency without compromising performance or adapting the model to real-time processing requirements. Additionally, investigating the model's robustness to diverse and complex scenes would be valuable for practical applications. Thorough analyses and extensive experiments showcase the superiority of our method over previous video inpainting approaches in terms of both qualitative and quantitative performance. However, acknowledging potential limitations and addressing them in future research will contribute to the continued advancement of video inpainting techniques.

ACKNOWLEDGMENT

(Taewan Kim and Jinwoo Kim are co-first authors.)

REFERENCES

[1] T. Shiratori, Y. Matsushita, X. Tang, and S. B. Kang, "Video completion by motion field transfer," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 411–418.

[2] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1150–1163, Jul. 2006.

[3] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Temporally coherent completion of dynamic video," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, Nov. 2016.

[4] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Free-form video inpainting with 3D gated convolution and temporal PatchGAN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9065–9074.

[5] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5792–5801.

[6] C. Wang, H. Huang, X. Han, and J. Wang, "Video inpainting by jointly learning temporal structure and spatial details," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5232–5239.

[7] S. Lee, S. W. Oh, D. Won, and S. J. Kim, "Copy-and-paste networks for deep video inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4413–4421.

[8] S. W. Oh, S. Lee, J.-Y. Lee, and S. J. Kim, "Onion-peel networks for deep video completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4403–4412.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[12] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[14] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.

[15] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," 2021, *arXiv:2102.07074*.

[16] C. Zheng, T.-J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," 2021, *arXiv:2104.00845*.

[17] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," 2021, *arXiv:2104.00272*.

[18] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[19] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," 2020, *arXiv:2004.11886*.

[20] I. Bello, "LambdaNetworks: Modeling long-range interactions without attention," 2021, *arXiv:2102.08602*.

[21] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4479–4488.

[22] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask in painting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2149–2159.

[23] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 417–424.

[24] S. Esedoglu and J. Shen, "Digital inpainting based on the Mumford–Shah–Euler image model," *Eur. J. Appl. Math.*, vol. 13, no. 4, pp. 353–370, Aug. 2002.

[25] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1200–1211, 2001.

[26] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[27] A. Newson, A. Almansa, Y. Gousseau, and P. Pérez, "Non-local patch-based image inpainting," *Image Process. Online*, vol. 7, pp. 373–385, Dec. 2017.

[28] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014.

[29] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.

[30] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1486–1494.

[31] M.-C. Sagong, Y.-G. Shin, S.-W. Kim, S. Park, and S.-J. Ko, "PEPSI: Fast image inpainting with parallel decoding network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11360–11368.

[32] Y.-G. Shin, M.-C. Sagong, Y.-J. Yeo, S.-W. Kim, and S.-J. Ko, "PEPSI++: Fast and lightweight network for image inpainting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 252–265, Jan. 2021.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[34] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*.

[35] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 181–190.

[36] H. Wu and J. Zhou, "IID-net: Image inpainting detection network via neural architecture search and attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1172–1185, Mar. 2022.

[37] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[38] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7760–7768.

[39] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[40] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 463–476, Mar. 2007.
[41] D. Jin and X. Bai, "Patch-sparsity-based image inpainting through a facet deduced directional derivative," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1310–1324, May 2019.
[42] R. Xu, X. Li, B. Zhou, and C. C. Loy, "Deep flow-guided video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3723–3732.
[43] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf, "Flow-edge guided video completion," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 713–729.
[44] C. Wang, X. Chen, S. Min, J. Wang, and Z.-J. Zha, "Structure-guided deep video inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 2953–2965, Aug. 2021.
[45] Z. Xu, Q. Zhang, Z. Cao, and C. Xiao, "Video background completion using motion-guided pixel assignment optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 8, pp. 1393–1406, Aug. 2016.
[46] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial–temporal transformations for video inpainting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 528–543.
[47] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li, "Decoupled spatial–temporal transformer for video inpainting," 2021, *arXiv:2104.06637*.
[48] H. J. Nussbaumer, "The fast Fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Berlin, Germany: Springer, 1981, pp. 80–111.
[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[50] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, 2018, pp. 85–100.
[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[52] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[53] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.
[55] Y.-L. Chang, Z. Yu Liu, K.-Y. Lee, and W. Hsu, "Learnable gated temporal shift module for deep video inpainting," 2019, *arXiv:1907.01131*.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[57] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
[58] K. Zhang, J. Fu, and D. Liu, "Flow-guided transformer for video inpainting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 74–90.
[59] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "YouTube-VOS: A large-scale video object segmentation benchmark," 2018, *arXiv:1809.03327*.
[60] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset, "The 2018 Davis challenge on video object segmentation," 2018, *arXiv:1803.00557*.
[61] X. Zou, L. Yang, D. Liu, and Y. J. Lee, "Progressive temporal feature alignment network for video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16448–16457.
[62] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," 2018, *arXiv:1808.06601*.
[63] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.

TAEWAN KIM received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively. From 2015 to 2021, he was with the Vision AI Laboratory, SK Telecom, Seoul. In 2022, he joined the Faculty of Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, where he is an Assistant Professor. His research interests include computer vision and machine learning, including continual and online learning. He has participated in the international activities, as the Industry and Exhibition Committee Chair of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seoul, in 2024.

JINWOO KIM received the B.S. degree in electrical and electronic engineering from Hongik University, Seoul, South Korea, in 2016. He is currently pursuing the combined M.S. and Ph.D. degrees with the Multi-Dimensional Insight Laboratory, Yonsei University. His current research interests include low-level computer vision, 3-D reconstruction, perceptual image, video processing, and generative models for image, video, motion, and audio.

HEESEOK OH received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2010, 2012, and 2017, respectively. He was a Senior Engineer with Samsung Electronics, Seoul. From 2017 to 2022, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. He is currently an Assistant Professor with the Department of Applied Artificial Intelligence, Hansung University, Seoul. His research interests include 2-D/3-D vision based on a human visual systems, extended reality, and generative models dealing with multimodal alignment.

JIWOO KANG (Member, IEEE) received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2011, and the Ph.D. degree from the Integrated Ph.D. Program in Electrical and Electronic Engineering, Yonsei University, in 2019. He was a Researcher with Yonsei University, from September 2019 to November 2020; and a Research Professor with the Y-BASE R&E Institute, Yonsei University, from December 2020 to February 2022. He has been an Assistant Professor with the Division of Artificial Intelligence Engineering, Sookmyung Women's University, Seoul, since 2022. His research interests include computer graphics, computer vision, and image processing.