Decouple and Couple: Exploiting Prior Knowledge for Visible Video Watermark Removal

Junye Chen, Chaowei Fang, *Member, IEEE*, Jichang Li, Yicheng Leng, and Guanbin Li, *Member, IEEE*

***Abstract*— This paper aims to restore original background images in watermarked videos, overcoming challenges posed by traditional approaches that fail to handle the temporal dynamics and diverse watermark characteristics effectively. Our method introduces a unique framework that first "decouples" the extraction of prior knowledge—such as common-sense knowledge and residual background details—from the temporal modeling process, allowing for independent handling of background restoration and temporal consistency. Subsequently, it "couples" these extracted features by integrating them into the temporal modeling backbone of a video inpainting (VI) framework. This integration is facilitated by a specialized module, which includes an intrinsic background image prediction sub-module and a dual-branch frame embedding module, designed to reduce watermark interference and enhance the application of prior knowledge. Moreover, a frame-adaptive feature selection module dynamically adjusts the extraction of prior features based on the corruption level of each frame, ensuring their effective incorporation into the temporal processing. Extensive experiments on YouTube-VOS and DAVIS datasets validate our method's efficiency in watermark removal and background restoration, showing significant improvement over state-of-the-art techniques in visible image watermark removal, video restoration, and video inpainting.**

***Index Terms*— Visible video watermark removal, prior knowledge extraction, temporal modeling.**

I. Introduction

VISIBLE watermarking serves as a critical tool for copyright protection, enabling creators to safeguard their intellectual property [4], [5]. Developing visible watermark removal techniques is valuable for evaluating the robustness and security of watermarks [1], [5]. This paper addresses the pressing need for reliable and robust visible video watermark removal strategies, a relatively under-explored area despite the proliferation of videos on the Internet.

The task of removing visible watermarks entails restoring the background image from the watermark-distorted content, posing a challenge compounded by the unpredictable variety of watermark characteristics such as size, orientation, and opacity. Conventional visual obstruction patterns such as haze [6], snow [7], or rains [8]) usually have regular underlying texture patterns which are obviously distinct to the background content. However, watermarks have diversified irregular patterns and may cause severe distortion to the background content. These issues bring unique obstacles to the task of visible watermark removal.

Although significant strides have been made in image watermark removal [1], [4], [5], [9], [10], [11], translating these successes into video content is hampered by the inability of existing methods to account for temporal information, a crucial element for maintaining consistency across frames. Video restoration (VR) methods [2], [12], [13] effectively model temporal relationships between frames but are designed for conventional obstruction patterns. Due to the diversity and high obstruction level of visible watermarks, they have limited performance in removing complex watermarks (see Figure 7). Video inpainting (VI), considered the most promising option, also falls short in precise content recovery. While VI thoroughly removes content in corrupted regions and generates visually plausible and coherent content [14], it often compromises the accuracy of background restoration, particularly in cases of extensive watermark coverage or intricate background textures (see Figure 7). Moreover, entirely removing content in the watermarked region leads to excessive loss of background information.

To address these issues, we revisit two key aspects of visible video watermark removal: temporal coherence across frames and thorough watermark removal with high-quality background content recovery. For temporal coherence, most methods utilize attention-based mechanisms [15], [16], [17], [18] or deformable convolutions [19], [20] to aggregate and propagate features across the time series, thereby modeling temporal information [2]. The latter aspect requires effectively erasing watermarks of varying opacity and generating realistic content in the occluded regions.

In this work, we introduce a novel "decouple and couple" framework to address the challenge of visible video watermark removal. This framework seamlessly integrates frame-specific prior features into a temporal modeling backbone designed for temporal relation establishment. In particular, two types of prior features are considered: common-sense knowledge which is critical for repairing areas severely corrupted by watermarks and intrinsic background residuals valuable for precisely

Received 17 June 2024; revised 19 December 2024; accepted 15 January 2025. Date of publication 4 February 2025; date of current version 18 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62376206 and Grant 62322608 and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010255. The associate editor coordinating the review of this article and approving it for publication was Prof. Raja Giryes. *(Corresponding authors: Guanbin Li; Chaowei Fang.)*

Junye Chen is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: chenjy856@mail2.sysu.edu.cn).

Chaowei Fang is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: chaoweifang@outlook.com).

Jichang Li is with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: li.jichang@pcl.ac.cn).

Yicheng Leng is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: 117010115@link.cuhk.edu.cn).

Guanbin Li is with the School of Computer Science and Engineering and Guangdong Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TIP.2025.3534033

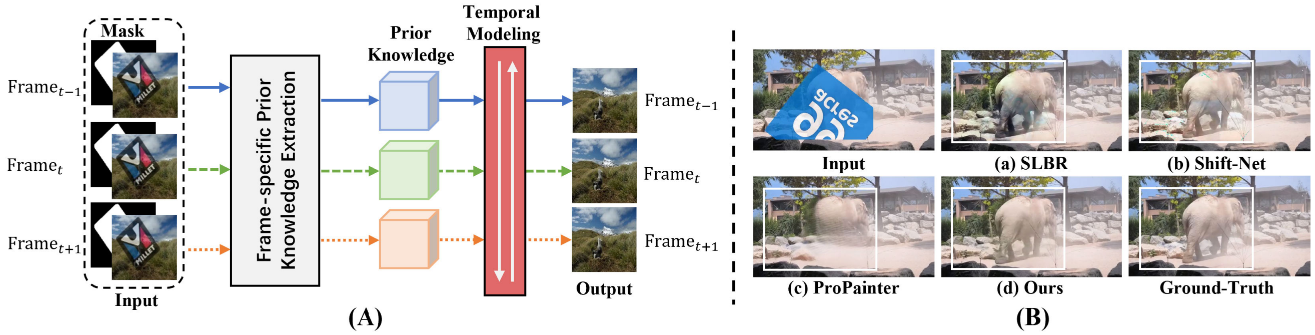

Fig. 1. (A) Visible video watermark removal involves eliminating the watermark according to the location of masks and restoring the original background of the frames. Our approach entails extracting prior knowledge for restoration from each frame and propagating this information temporally to recover the background. (B) Visual comparisons from different representative methods: (a) Image Watermark Removal [1], (b) Video Restoration [2], (c) Video Inpainting [3], our DECO outperforms other types of methods in watermark removal and content recovery.

restoring the original content of corrupted areas. To achieve this target, a novel frame-specific prior feature extraction (FPFE) module is introduced to decouple the extraction of prior information from the temporal modeling process, facilitating the frame restoration ability under large-area watermark occlusion and high-opacity watermarking. To be specific, this module begins with an intrinsic background image prediction (IBIP) sub-module to reduce watermark interference, and adopts a dual-branch frame embedding (DBFE) sub-module for effective prior feature extraction. The DBFE sub-module leverages a quantized knowledge retrieval branch and a straightforward translation branch to explore common-sense knowledge and intrinsic background residuals, respectively. Furthermore, we propose a frame-adaptive feature selection module that effectively couples the extracted prior features with the temporal modeling process. We conduct extensive experiments on YouTube-VOS [21] and DAVIS [22] datasets corrupted by diverse watermarks. The results demonstrate that our method achieves significantly better performance than existing techniques in the field of visible image watermark removal, video restoration, and video inpainting.

The main contributions of this paper are summarized as follows:

- We introduce a novel **DE**couple and **CO**uple (**DECO**) framework that extracts frame-wise prior features and adaptively integrates them into a temporal modeling backbone for visible video watermark removal.
- We introduce a new prior feature extraction module, which comprises an intrinsic background image prediction and a dual-branch frame embedding sub-modules, effectively utilizing common-sense knowledge and intrinsic background residuals.
- Extensive experiments on YouTube-VOS [21] and DAVIS [22] demonstrate our method's effectiveness and adaptability. DECO significantly outperforms existing techniques in visible image watermark removal, video restoration, and video inpainting.

II. Related Work

A. Visual Watermark Removal

In visual media, pasting visible watermarks into images is a widely used method for copyright protection. Conversely, visible watermark removal is regarded as an adversarial technology that has attracted significant interest. Prior approaches for visible watermark removal, such as those in [1], [4], [5], [9], and [10], mainly utilize multi-task learning networks to locate visible watermarks and restore the visual background simultaneously. For instance, Liang et al. [1] propose watermark localization by predicting corresponding masks and leveraging cross-level features to enhance output quality, achieving high-quality watermark-free background restoration. In comparison, Cun and Pun [4] propose adopting a two-stage "split-then-refine" framework for watermark localization and background restoration. Additionally, Sun et al. [5] distinguish watermark and background semantic embeddings in high-dimensional space, achieving state-of-the-art performance. However, these methods tailored for image watermark removal lack the capability to model temporal information, crucial for addressing the temporal consistency required in video content. Video watermark removal necessitates frame-by-frame restoration while maintaining temporal consistency for aesthetically pleasing results. In contrast to established image techniques, video watermark removal is seldom studied in academia. Our work aims to bridge this gap by addressing the unique challenges of removing visible watermarks from videos.

B. Video Restoration

Video restoration focuses on restoring high-quality content from videos affected by various noises, such as deraining [8], snow removal [7], denoising [13], and dehazing [6]. These "noises" generally cause limited background corruption or significantly differ semantically from the background. Thus, existing restoration methods model and remove noise based on these semantic differences or propagate features from aligned adjacent frames. For instance, Wang et al. [8] separate rain from backgrounds using mutual exclusion, while Xu et al. [6] and Chen et al. [7] utilize atmospheric scattering models and their variants for background restoration. Maggioni et al. [13] enhance corrupted pixels by aggregating features from adjacent frames. In the VVWR task, watermarks in video frames pose a more significant challenge due to their complexity and potential semantic similarity to the background, causing more severe content corruption and greater recovery challenges.

C. Video Inpainting

Video inpainting aims to fill missing areas in video sequences with spatially and temporally coherent content. It typically removes objects using a binary mask, while effective for object removal, which results in significant content loss. This poses challenges for video inpainting's feature propagation and alignment capabilities, impacting visual quality. Ren et al. [23] propose to encode frame features into discrete latent codes and use a transformer to model temporal dependencies and predict the indices of discrete codes in the missing regions across time series. Zhang et al. [24] leverage optical flow completed by a flow completion network to guide the transformer to fill missing content. Li et al. [14] introduce an end-to-end framework for accurate flow completion, enhancing feature propagation and content repairing. Zhou et al. [3] further advance feature propagation and alignment both spatially and temporally. However, these methods [3], [14], [24] are often misled by inaccurate completed optical flow and struggle with generating large-area content, resulting in sub-optimal outcomes with blurred backgrounds and unpleasant artifacts. Lee et al. [25] propose a semantic-aware transformer based on a mixture-of-experts scheme, achieving high-quality content recovery in large-area occlusions by leveraging semantic information selection within inputs. In this work, we harness prior knowledge from watermark-free images for feature restoration and collaborate with a temporal video inpainting backbone for feature alignment and propagation from adjacent frames, significantly improving the performance of video watermark removal.

III. The Proposed Method

A. Preliminaries

1) Problem Formulation: Visible video watermark removal (VVWR) aims to restore the original video from the input watermarked video corrupted by visible watermarks. A triplet dataset $< \mathbb{X}, \mathbb{M}^*, \mathbb{X}^* >$ is used to develop our method. Here, $\mathbb{X} = \{\mathbf{X}_t\}_{t=1}^{T}$ represents the watermarked video with T frames, $\mathbb{M}^* = \{\mathbf{M}_t^*\}_{t=1}^{T}$ represents the corresponding masks of coarsely annotated watermarks, and $\mathbb{X}^* = \{\mathbf{X}_t^*\}_{t=1}^{T}$ represents the original background (ground-truth) video. Each frame $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$ denotes an image with dimensions of height H, width W, and 3 channels. $\mathbf{M}_t^* \in \mathbb{R}^{H\times W\times 1}$ denotes a binary mask with dimensions of height H, width W, and 1 channel. In form, the goal of VVWR is to convert a watermarked video $\mathbb{X}$ into a watermark-free video $\mathbb{Y} = \{\mathbf{Y}_t\}_{t=1}^{T}$ under the guidance of $\mathbb{M}^*$.

2) Generation of Watermarked Video: Following the definition of watermarked image in [10], the watermarked video frame can be generated by blending the watermark image $\mathbf{W}_t \in \mathbb{R}^{H\times W\times 3}$ with the original background image $\mathbf{X}_t^*$:

$$\mathbf{X}_t = (1 - \alpha\mathbf{M}_t) \odot \mathbf{X}_t^* + \alpha\mathbf{M}_t \odot \mathbf{W}_t, \tag{1}$$

where $\mathbf{M}_t \in \mathbb{R}^{H\times W\times 1}$ represents the binary mask of watermark, $\alpha \in (0, 1]$ is the opacity level of the watermark, and $\odot$ denotes element-wise multiplication. The term highlighted in blue of Eqn. (1) is referred to as the intrinsic background image. To simplify, we denote $\mathbf{G}_t^{int} = (1 - \alpha\mathbf{M}_t) \odot \mathbf{X}_t^*$. In Section III-B, we utilize $\mathbf{G}_t^{int}$ as the ground truth to supervise our model in learning to eliminate the interference of the watermark.

3) Overview: In this paper, we introduce the decouple and couple framework to address the challenge of visible video watermark removal. The term "decouple" refers to our approach of separating the extraction of prior features, such as common-sense knowledge and residual background information, from the temporal modeling process. This separation is facilitated by the Frame-specific Prior Feature Extraction (FPFE) module, which includes the Intrinsic Background Image Prediction (IBIP) sub-module and the Dual-Branch Frame Embedding (DBFE) sub-module, allowing for independent handling of feature extraction and restoration. The term "couple" denotes the subsequent integration of these extracted prior features into the temporal modeling backbone, ensuring temporal coherence. This integration is managed by the Frame-adaptive Feature Selection Module (FFSM), which adaptively incorporates the prior features based on the corruption level indicated by the IBIP features, thus maintaining temporal consistency. Figure 2 provides a comprehensive illustration of our proposed DECO framework.

B. Frame-Specific Prior Feature Extraction Module

This section details the Frame-specific Prior Feature Extraction (FPFE) module that harnesses prior information for background restoration in video frames. As illustrated in Figure 2, FPFE is composed of two key parts: the Intrinsic Background Image Prediction (IBIP) sub-module, and the Dual-Branch Frame Embedding (DBFE) sub-module. IBIP removes watermark interference and thus obtains the intrinsic background of corrupted regions in each video frame. Then, DBFE restores the corrupted regions in the intrinsic background image, comprising knowledge retrieval (KR) and straightforward translation (ST) branches. The KR branch retrieves common-sense knowledge extracted from natural images for feature reconstruction in the corrupted regions, whereas the ST branch focuses on directly restoring residual features. DBFE then merges the output features of both branches through a corruption-aware fusion component (CFC), optimizing restoration based on the corruption degree. Such a two-stage paradigm can effectively remove the watermark and leverage the common-sense knowledge and residual features to restore the frames.

Intrinsic Background Image Prediction (IBIP): From coarse to fine is a commonly used paradigm in watermark removal [1], [4], [5], which reduces the learning difficulty at each stage. Inspired by them, we propose IBIP with a U-Net-like [26] structure to mitigate the watermark interference by learning to predict the intrinsic background image at the first stage. This IBIP sub-module comprises an encoder $\mathrm{E_{IBIP}}$ and a decoder $\mathrm{D_{IBIP}}$, each consisting of 4 and 3 residual blocks [27], respectively. Furthermore, neural networks excel at fitting residual noise rather than restoring images [28]. In this work, we regard watermarks as noise in the images and employ a long skip connection to encourage the network to learn watermark noise directly.

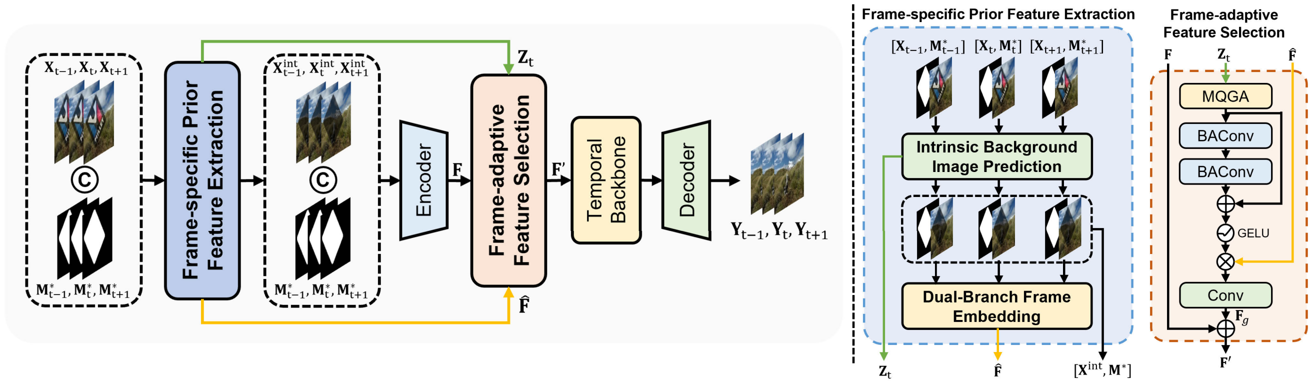
X_{t-1}, X_t, X_{t+1}
©
$M^*_{t-1}, M^*_t, M^*_{t+1}$
Frame-specific Prior Feature Extraction
$X^{int}_{t-1}, X^{int}_t, X^{int}_{t+1}$
Z_t
Encoder
F
Frame-adaptive Feature Selection
F′
Temporal Backbone
Decoder
$\hat{F}$
Y_{t-1}, Y_t, Y_{t+1}
Frame-specific Prior Feature Extraction
$[X_{t-1}, M^*_{t-1}]$ $[X_t, M^*_t]$ $[X_{t+1}, M^*_{t+1}]$
Intrinsic Background Image Prediction
Dual-Branch Frame Embedding
$[X^{int}, M^*]$
Frame-adaptive Feature Selection
MQGA
BAConv
BAConv
GELU
Conv
F_g

Fig. 2. The overview of the proposed method **DECO**. In this work, we first present a Frame-specific Prior Feature Extraction module (FPFE), leveraging common-sense knowledge and maximizing the exploitation of residual background content. Then, we present the Frame-adaptive Feature Selection Module (FFSM) designed to incorporate better the FPFE, associated with a temporal backbone for temporal dependency modeling. We detail the sub-modules of FPFE in Fig. 3.

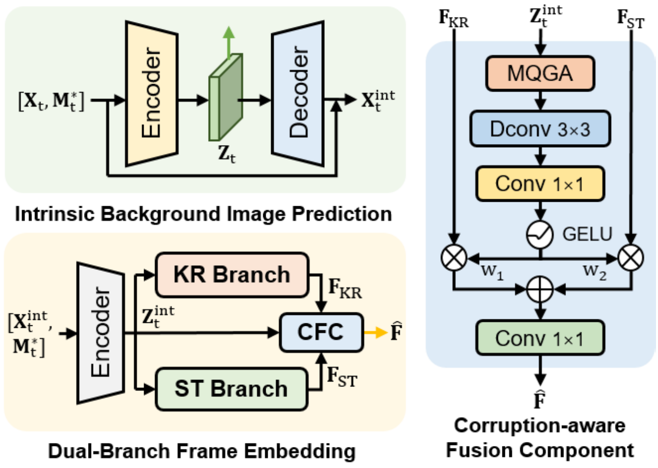

Fig. 3. The illustration of our IBIP and DBFE sub-module of FPFE. "CFC" represents the corruption-aware fusion component. The skip-connection of IBIP is hidden for visual comfort. The details of the KR branch are further elaborated in Fig. 4.

Specifically, let $\mathbf{X}_t^{int} \in \mathbb{R}^{H\times W\times 3}$ represent the intrinsic background image corresponding to the input image $\mathbf{X}_t$, which can be computed as follows:

$$\mathbf{X}_t^{int} = \mathrm{D}_{\mathrm{IBIP}}(\mathrm{E}_{\mathrm{IBIP}}([\mathbf{X}_t, \mathbf{M}_t^*])) + \mathbf{X}_t. \tag{2}$$

For enabling IBIP to capture the intrinsic background from the watermarked image $\mathbf{X}_t$, we utilize $\mathbf{G}_t^{int}$ (defined in Eqn. 1) as the ground truth to supervise the optimization of the model through the following training loss:

$$\mathcal{L}_{IBIP} = \frac{||\mathbf{M}_t^* \odot (\mathbf{X}_t^{int} - \mathbf{G}_t^{int})||_1}{||\mathbf{M}_t^*||_1} + \frac{||(1-\mathbf{M}_t^*) \odot (\mathbf{X}_t^{int} - \mathbf{G}_t^{int})||_1}{||1-\mathbf{M}_t^*||_1}, \tag{3}$$

where $||\cdot||_1$ refers to the L_1 norm function.

Dual-Branch Frame Embedding (DBFE): In the second stage, we propose DBFE to restore the corrupted regions in the intrinsic background image. To be specific, we first feed the intrinsic background image $\mathbf{X}_t^{int}$ with its corresponding mask

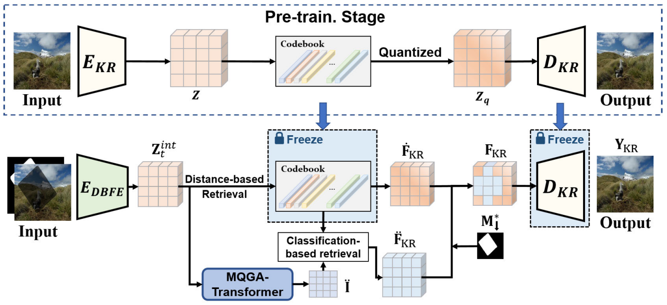

Fig. 4. The architecture of our KR branch in DBFE. The first row pre-trains the codebook and decoder of the network by a reconstruction task, which is then frozen to restore corruption. The second row restores the background by retrieving knowledge from the pre-trained codebook using different ways in the intact and corrupted regions.

$\mathbf{M}_t^*$ into the DBFE encoder $\mathrm{E}_{\mathrm{DBFE}}$, yielding the feature map $\mathbf{Z}_t^{int} = \mathrm{E}_{\mathrm{DBFE}}([\mathbf{X}_t^{int}, \mathbf{M}_t^*]) \in \mathbb{R}^{h\times w\times c}$ that is subsequently fed into the KR and ST branches and the corruption-aware fusion component for feature restoration.

1) The KR Branch: Inspired by the success of the super-resolution task demonstrated in [29], which proves that pre-trained codebooks and decoders can extract and store prior knowledge from natural images, we utilize two image datasets to pre-train a VQ-VAE-like [30] network with an efficient codebook [31]. The codebook, denoted as $\mathbb{C} = \{\hat{\mathbf{c}}_k\}_{k=1}^{K} \subset \mathbb{R}^{c'}$, is a representation dictionary comprising K prototype embeddings, which is trained to learn discrete latent representations for image generation tasks [32]. Thus, leveraging extensive and diverse scenarios in image datasets, the pre-trained VQ-VAE-like network can learn rich prior knowledge from various scenarios for corruption reconstruction. The process of pre-training stage of this network is shown in Figure 4.

We found that directly mapping features from corrupted regions to their corresponding intact embeddings in the codebook for reconstruction presents optimization challenges. Therefore, we deliberately designed two retrieval strategies to maximize the diversity and relevance of prior features for the intact and corrupted regions. Specifically, we retrieve the embeddings from the codebook $\mathbb{C}$ in two ways: 1) Distance-based retrieval for intact regions, and 2) Classification-based retrieval for corrupted regions. In intact regions, following [30], for each spatial position on the feature map $\mathbf{Z}_t^{int}$, we retrieve the nearest embedding $\hat{\mathbf{c}}_k \in \mathbb{C}$ corresponding to the feature vector $\mathrm{Z}_t^{(ij)} \in \mathbf{Z}_t^{int}$, thus obtain the embedding map $\dot{\mathbf{F}}_{\mathrm{KR}}$ as follows,

$$\dot{\mathbf{F}}_{\mathrm{KR}} := \left(\arg\min_{\hat{\mathbf{c}}_k \in \mathbb{C}} ||\mathrm{Z}_t^{(ij)} - \hat{\mathbf{c}}_k||\right) \in \mathbb{R}^{h\times w\times c'}. \tag{4}$$

This strategy of knowledge retrieval effectively reduces information loss in intact regions and ensures that the retrieved features are contextually aligned with the target frame, contributing to the precise reconstruction of the background. On the other hand, the classification-based retrieval approach utilizes categorical priors that are representative of broader common-sense knowledge. Specifically, to overcome the optimization challenges in corrupted regions, we simplify the feature-mapping task into a classification task. Inspired by [33], we employ a mask-query guided attention transformer (MQGA-Transformer) to capture the correlations between corrupted and intact regions and then predict the corresponding embedding indices of the corrupted regions, reducing the learning difficulty. Specifically, given the inputs $\mathbf{Z}_t^{int}$ and the downsampled mask $\mathbf{M}_\downarrow^* \in \mathbb{R}^{h\times w\times 1}$, we define the specific MQGA function $\mathrm{H}_{\mathrm{MQGA}}$ as follows,

$$\mathrm{H}_{\mathrm{MQGA}}(\mathbf{Z}_t^{int}, \mathbf{M}_\downarrow^*) = softmax\left(\frac{(\mathbf{Q}\odot\mathbf{M}_\downarrow^*)\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{5}$$

where $\mathbf{Q} = \mathbf{Z}_t^{int}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{Z}_t^{int}\mathbf{W}^K$, and $\mathbf{V} = \mathbf{Z}_t^{int}\mathbf{W}^V$, with $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ denoting linear projections, and $d_k \in \mathbb{R}$ represents the feature dimension. We construct the MQGA-Transformer block $\mathcal{B}_{\mathrm{MQGA}}$ by replacing the vanilla self-attention mechanism with MQGA in the standard transformer block [34]. Consequently, the MQGA-Transformer $\mathcal{T}_{\mathrm{MQGA}}$ is formed by stacking multiple MQGA-Transformer blocks $\mathcal{B}_{\mathrm{MQGA}}$, which predicts the index of embedding at each position by the softmax function. The predicted index map $\ddot{\mathbf{I}} \in \mathbb{R}^{h\times w\times 1}$ and its corresponding embedding map $\ddot{\mathbf{F}}_{\mathrm{KR}} \in \mathbb{R}^{h\times w\times c'}$ retrieved from the codebook $\mathbb{C}$ can be computed as,

$$\ddot{\mathbf{I}} = \mathcal{T}_{\mathrm{MQGA}}(\mathbf{Z}_t^{int}, \mathbf{M}_\downarrow^*), \quad \ddot{\mathbf{F}}_{\mathrm{KR}} = \mathrm{query}(\ddot{\mathbf{I}}), \tag{6}$$

where query$(\cdot)$ denotes the query function that retrieves the embedding map $\ddot{\mathbf{F}}_{\mathrm{KR}}$ from codebook $\mathbb{C}$ according to the index map $\ddot{\mathbf{I}}$. This approach enables the model to leverage general patterns and structures, which are particularly beneficial in scenarios where significant portions of the background are occluded or damaged by watermarks. Finally, we obtain the embedding map $\mathbf{F}_{\mathrm{KR}} \in \mathbb{R}^{h\times w\times c'}$ by combining the two embedding maps above, and feed it into the decoder D_{KR} to reconstruct the background image $\mathbf{Y}_{\mathrm{KR}}$,

$$\mathbf{F}_{\mathrm{KR}} = \dot{\mathbf{F}}_{\mathrm{KR}} \odot (1 - \mathbf{M}_\downarrow^*) + \ddot{\mathbf{F}}_{\mathrm{KR}} \odot \mathbf{M}_\downarrow^*, \tag{7}$$

$$\mathbf{Y}_{\mathrm{KR}} = \mathrm{D}_{\mathrm{KR}}(\mathbf{F}_{\mathrm{KR}}), \tag{8}$$

where the subscript t of $\mathbf{Y}_{\mathrm{KR}}$ is hidden for brevity. By integrating these two complementary strategies, the KR branch ensures both specificity and generalizability in feature retrieval, thereby enhancing the robustness of the restoration process. The following loss function is exploited to optimize the KR branch:

$$\begin{aligned}\mathcal{L}_{KR} &= ||\mathbf{X}_t^* - \mathbf{Y}_{\mathrm{KR}}||_1 + \lambda_{\mathcal{T}_{\mathrm{MQGA}}}\mathcal{L}_{\mathcal{T}_{\mathrm{MQGA}}} + \lambda_{adv}\mathcal{L}_{adv},\\ \mathcal{L}_{\mathcal{T}_{\mathrm{MQGA}}} &= \sum_{(i,j)} -\mathbf{I}_{GT}^{(ij)}\log(\ddot{\mathbf{I}}^{(ij)}),\\ \mathcal{L}_{adv} &= \sum -\mathbb{E}[\mathrm{D}_{\mathrm{P}}(\mathbf{Y}_{\mathrm{KR}})],\end{aligned} \tag{9}$$

where $\mathbf{X}_t^*$ represents the ground truth of the background image. Additionally, the ground truth of index map $\mathbf{I}_{GT}$ is obtained by encoding $\mathbf{X}_t^*$ with E_{KR} and utilizing its feature map to retrieve the embeddings from the pre-trained codebook $\mathbb{C}$, which is used to supervise the MQGA-Transformer. Similar to [35], $\mathcal{L}_{adv}$ is an adversarial loss measured by a PatchGAN [35] discriminator D_{P} with hinge loss to generate realistic output. $\lambda_{\mathcal{T}_{\mathrm{MQGA}}}$ and λ_{adv} are balanced weights that are set to 0.5 and 0.01, respectively.

2) The ST Branch: For residual background features restoration, we employ Restormer blocks [36] to construct the straightforward translation (ST) branch within our DBFE sub-module. Benefiting from the remarkable performance of Restormer [36] in feature restoration, our ST branch is capable of effectively restoring residual background features. In this work, we construct the ST branch by stacking six Restormer blocks and denote it as $\mathcal{R}_{\mathrm{ST}}$. Then, upon inputting the encoded feature $\mathbf{Z}_t^{int}$ into the ST branch, the restored feature $\mathbf{F}_{\mathrm{ST}}$ can be derived as follows,

$$\mathbf{F}_{\mathrm{ST}} = \mathcal{R}_{\mathrm{ST}}(\mathbf{Z}_t^{int}). \tag{10}$$

This output feature $\mathbf{F}_{\mathrm{ST}}$ is then processed by a reconstruction decoder, denoted as D_{ST}, to produce the reconstructed image, denoted as $\mathbf{Y}_{\mathrm{ST}} = \mathrm{D}_{\mathrm{ST}}(\mathbf{F}_{\mathrm{ST}})$, where the subscript t of $\mathbf{Y}_{\mathrm{ST}}$ is also hidden for brevity. To supervise the reconstructed image $\mathbf{Y}_{\mathrm{ST}}$ with its corresponding ground truth $\mathbf{X}_t^*$, we employ the loss function $\mathcal{L}_{ST}$, which is represented as follows,

$$\begin{aligned}\mathcal{L}_{ST} &= ||\mathbf{X}_t^* - \mathbf{Y}_{\mathrm{ST}}||_1 + \lambda_{perc}\cdot\mathcal{L}_{perc},\\ \mathcal{L}_{perc} &= \sum_{v\in\{1,2,3\}} ||\phi_{\mathrm{VGG}}^v(\mathbf{G}) - \phi_{\mathrm{VGG}}^v(\mathbf{Y}_{\mathrm{ST}})||_1.\end{aligned} \tag{11}$$

Here, we enhance the visual quality of the output $\mathbf{Y}_{\mathrm{ST}}$ by the perceptual loss $\mathcal{L}_{perc}$ [37] with a balance weight λ_{perc} set to 0.25, where $\phi_{\mathrm{VGG}}^v(\cdot)$ indicates the activation map at the v-th layer of VGG16 [38].

3) The Corruption-Aware Fusion Component: Because watermarks with different opacity levels lead to different corruption degrees, it is necessary to tailor the restoration for different levels of corruption. For watermarks with low opacity, using features directly restored from the ST branch can avoid quantization loss. Conversely, direct translation may lead to incorrect feature recovery for severe corruption, which can be avoided by the features from KR branches. We model the correlation between corrupted and intact regions to predict the degree of corruption, thereby achieving the adaptive feature fusion from different branches to adapt to the watermarks with different opacity.

Specifically, given the input $\mathbf{Z}_t^{int}$, we leverage the MQGA function to model the correlation between corrupted and intact regions and generate the similarity map, then determine the adaptive fusion weights by refining the similarity map for merge features from the two branches. This process can be computed as follows:

$$\begin{aligned} \mathbf{S} &= \mathrm{H}_{\mathrm{MQGA}}(\mathbf{Z}_t^{int}, \mathbf{M}_{\downarrow}^{*}), \\ [\mathbf{w}_1, \mathbf{w}_2] &= Split(\mathrm{GELU}(\mathrm{Conv}_{1\times1}(\mathrm{Dconv}_{3\times3}(\mathbf{S})))), \end{aligned} \tag{12}$$

where $Split(\cdot)$ represents the splitting function along the channel dimension. Therefore, we can obtain the final fusion feature $\hat{\mathbf{F}}$ by the predicted fusion weights for reconstruction as follows,

$$\hat{\mathbf{F}} = \mathrm{Conv}_{1\times1}(\mathbf{F}_{\mathrm{KR}} \odot \mathbf{w}_1 + \mathbf{F}_{\mathrm{ST}} \odot \mathbf{w_2}). \tag{13}$$

Afterwards, we feed $\hat{\mathbf{F}}$ into the fusion reconstruction decoder $\mathrm{D_f}$, and then yield its reconstructed image, i.e. $\mathbf{Y}_{fusion} = \mathrm{D_f}(\hat{\mathbf{F}})$. Similarly, the reconstruction loss can be subsequently constructed to optimize this component as follows,

$$\mathcal{L}_{fusion} = ||\mathbf{X}_t^{*} - \mathbf{Y}_{fusion}||_1. \tag{14}$$

As stated above, we train the DBFE sub-module by composing the three loss functions $\mathcal{L}_{KR}$, $\mathcal{L}_{ST}$, and $\mathcal{L}_{fusion}$ of three parts in an end-to-end manner as follows,

$$\mathcal{L}_{DBFE} = \mathcal{L}_{KR} + \mathcal{L}_{ST} + \mathcal{L}_{fusion}. \tag{15}$$

C. The DECO Framework

In this section, we introduce the Frame-adaptive Feature Selection Module (FFSM) and the temporal modeling backbone, respectively, and then summarize the whole framework we proposed. We begin by presenting the FFSM tailored for enhancing the integration with the FPFE module. Next, we adopt MSVT [3], a leading transformer backbone for video inpainting, as our temporal modeling backbone to ensure temporal consistency due to its effective capabilities in temporal information aggregation and propagation. Finally, we derive the training objective of the whole DECO framework.

1) Frame-Adaptive Feature Selection Module (FFSM): Variations in watermark opacity significantly impact the performance of removing visible watermarks from videos. A naive idea is to grade the watermark opacity and then use adaptive model weights for different opacity. However, the diversity of watermarks and backgrounds makes directly grading watermark opacity a challenging task with large ambiguity. This is because watermarks of the same opacity level may have different effects on light-colored and dark-colored backgrounds. It can be argued that the varying degrees of corruption caused by different watermark opacities necessitate distinct prior features. Therefore, to address this, we develop a Frame-adaptive Feature Selection Module, designed to determine the amount of prior features to be applied for feature restoration. As shown in Figure 2, by considering the corrupted information extracted from the feature map in IBIP as an indicator to adjust the prior features, we can achieve customized restoration tailored to varying degrees of corruption.

Algorithm 1 The Training Procedure of DECO

```
Input: Watermarked video $\mathbb{X} = \{\mathbf{X}_t\}_{t=1}^{T}$, Coarse watermark masks $\mathbb{M}^{*} = \{\mathbf{M}_t^{*}\}_{t=1}^{T}$, Ground-truth video $\mathbb{X}^{*} = \{\mathbf{X}_t^{*}\}_{t=1}^{T}$
Output: Optimal DECO Model
Pretrain the KR branch with static images to acquire the frozen codebook $\mathbb{C}$ along with the decoder $\mathrm{D_{KR}}$.
Pretrain IBIP with static images using Eqn. (3) to obtain $\mathrm{E_{IBIP}}$ and $\mathrm{D_{IBIP}}$.
Pretrain DBFE with static images using Eqn. (15).
Freeze the IBIP and DBFE of FPFE.
while Not Converged do
    Optimize the whole model on $\mathbb{X}$, $\mathbb{M}^{*}$ and $\mathbb{X}^{*}$ using Eqn. (20) in an end-to-end manner.
end while
```

To be specific, we utilize the encoded feature $\mathbf{Z}_t = \mathrm{E_{IBIP}}([\mathbf{X}_t, \mathbf{M}_t^{*}]) \in \mathbb{R}^{h\times w\times c}$ from IBIP as input to capture the correlation $\mathbf{C}$ between the corrupted and intact regions. Then we refine the correlation to generate the gate weight as follows,

$$\mathbf{C} = \mathcal{B}_{\mathrm{MQGA}}(\mathbf{Z}_t, \mathbf{M}_{\downarrow}^{*}), \tag{16}$$

$$\mathbf{w}_g = \mathrm{GELU}(\mathrm{BAConv}(\mathrm{BAConv}(\mathbf{C})) + \mathbf{C}), \tag{17}$$

where $\mathrm{BAConv}(\cdot) = \mathrm{Conv}_{1\times1}(\mathrm{Dconv}_{3\times3}(\mathrm{GELU}(\mathrm{BN}(\cdot))))$, with $\mathrm{BN}(\cdot)$ denoting batch normalization [39] and $\mathcal{B}_{\mathrm{MQGA}}(\cdot)$ representing the MQGA-Transformer block. Thus, the tailored feature map $\mathbf{F}'$ is obtained by,

$$\mathbf{F}_g = \mathrm{Conv}_{1\times1}(\hat{\mathbf{F}} \odot \mathbf{w}_g), \tag{18}$$

$$\mathbf{F}' = \mathbf{F} + \mathbf{F}_g. \tag{19}$$

2) Temporal Backbone: Accurate temporal alignment is crucial for temporal dependency across time series. Given the inherent temporal redundancy in video sequences, we can leverage this redundancy by aligning inter-frame features to enhance visual quality. In this work, we adopt the MSVT backbone from [3] as our temporal backbone for temporal alignment in video processing. It aligns features within a sliding window sequence and utilizes mask-guided sparse attention to dissect frame sequences into sub-window tokens for local and global alignment. This comprehensive approach ensures spatial and temporal coherence, significantly enhancing video restoration. The excellent blend of computational efficiency and performance makes efficient feature aggregation and propagation possible.

3) Training Procedure for DECO: Upon constructing our full DECO, following [3], we use L_1 as the reconstruction loss and an adversarial loss [40] for the generation of high-quality and realistic content in training the temporal modeling backbone. To be specific, we supervise the output $\mathbb{Y} = \{\mathbf{Y}_t\}_{t=1}^{T}$ by the ground truth video $\mathbb{X}^{*} = \{\mathbf{X}_{\mathbf{t}}^{*}\}_{t=1}^{T}$, the full model can be optimized using the loss function $\mathcal{L}_{full}$ as follows,

$$\begin{aligned} \mathcal{L}_{full} &= \sum_{t=1}^{T} ||\mathbf{Y}_t - \mathbf{X}_t^{*}||_1 + \lambda_{adv}\mathcal{L}_{adv}, \\ \mathcal{L}_{adv} &= \sum_{t=1}^{T} -\mathbb{E}[\mathrm{D_{TP}}(\mathbf{Y}_t)], \end{aligned} \tag{20}$$

where λ_{adv} is set to 0.01 and $\mathrm{D_{TP}}$ denotes the T-PatchGAN [40] discriminator. The whole training procedure of DECO is described in Algorithm 1.

TABLE I

QUANTITATIVE COMPARISONS BETWEEN OUR PROPOSED DECO AND PREVIOUS SOTA METHODS ON YOUTUBE-VOS AND DAVIS. HERE, ↑ DENOTES HIGHER IS BETTER AND ↓ INDICATES LOWER IS BETTER. THE RESULTS WITH THE BEST PERFORMANCE ARE MARKED IN **BOLD**

Methods	YouTube-VOS				DAVIS			
	PSNR↑	SSIM↑	RMSE$_w$↓	VFID↓	PSNR↑	SSIM↑	RMSE$_w$↓	VFID↓
Video Restoration								
RVRT [12]	19.80	0.7229	31.914	0.256	20.37	0.7258	31.328	0.967
EMVD [13]	19.42	0.7157	30.609	0.268	20.06	0.7251	29.175	0.734
Shift-Net [2]	31.71	0.9632	6.771	0.107	30.56	0.9634	7.532	0.176
Video Inpainting								
FuseFormer [41]	22.87	0.7735	21.000	0.224	22.21	0.7468	21.474	0.695
E^{2}FGVI [14]	26.08	0.8576	15.602	0.139	24.56	0.8349	16.778	0.503
ProPainter [3]	27.21	0.8781	14.125	0.118	25.77	0.8634	14.986	0.448
DECO (Ours)	**35.38**	**0.9803**	**5.113**	**0.050**	**33.34**	**0.9777**	**6.028**	**0.117**

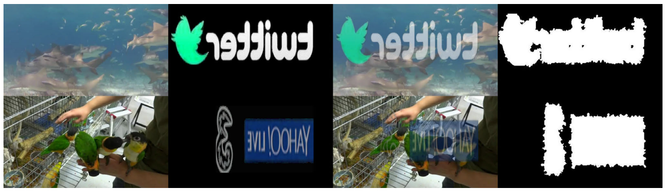

Fig. 5. Two examples of our dataset. From left to right are the original background, watermark, watermarked image, and the corresponding coarse mask.

TABLE II

STATISTICAL DIFFERENCES IN EVALUATION METRICS AND DATASET ATTRIBUTES W.R.T. WATERMARKS ON BOTH ILAW AND CLWD. THE LEFT HALF DISPLAYS THE AVERAGE METRICS OBTAINED FROM DIRECT COMPARISONS BETWEEN THE GIVEN WATERMARKED IMAGES AND THEIR ORIGINAL BACKGROUND IMAGES FOR EACH DATASET. "AP" AND "OPACITY" IN THE RIGHT HALF REPRESENT THE AVERAGE PROPORTION OF WATERMARK AREAS ACROSS ALL IMAGES AND THE AVERAGE OPACITY LEVEL OF THESE WATERMARKS, RESPECTIVELY

Dataset	PSNR↑	SSIM↑	RMSE$_w$↓	RMSE↓	AP	Opacity
ILAW	13.66	0.714	57.78	63.07	0.557	0.665
CLWD [10]	29.49	0.949	11.21	40.93	0.107	0.547

IV. EXPERIMENTS

A. Experimental Setups

1) Datasets: To empower IBIP capable of erasing watermark patterns from the backgrounds and enrich DBFE with extensive and diverse prior knowledge from extensive scenarios, we utilized two image datasets to train the IBIP and DBFE sub-modules. CLWD [10] is a popular watermark image dataset, comprising 60,000 images with 160 different colored watermarks for training, and 10,000 images with 40 watermarks for testing. Further, we propose an image dataset called Images with Large-Area Watermarks (ILAW). The training set of ILAW comprises 60,000 images of size 256 × 256, featuring 1,087 distinct colored watermarks, while its testing set consists of 10,000 images sized 512 × 512, incorporating 160 watermarks. The background images are sourced from the Places365 Challenge dataset [42], whereas the watermarks are collected from the Internet. The opacity level of the watermarks is set within the range of [0.3, 1.0]. We generate the watermarked image by blending the background image with a watermark under the guidance of a binary mask and the opacity of the watermark. Compared to CLWD, ILAW offers a greater variety of watermark patterns and larger sizes, enhancing the model's generalization capability to handle different types of watermarks. More details of ILAW are shown in Table II, which suggests that ILAW is a more difficult dataset.

To assess the effectiveness of DECO in addressing the task of visible video watermark removal (VVWR) task, we tested it on the YouTube-VOS [21] and DAVIS [22] datasets. YouTube-VOS includes 3,471 training clips, 474 for validation, and 508 for testing; DAVIS contains 60 training and 90 testing videos. Following [3] and [14], we trained our DECO model using 3,471 videos from the YouTube-VOS training set as the original background videos. Additionally, we incorporated 1,087 watermarks from the ILAW training set, which include both simple single-color designs and complex nested patterns. Each watermark was randomly selected and applied with a random opacity within [0.1, 1.0] to blend with the original background video. We randomly generate stationary and moving watermarked videos with a probability of 50% and the corresponding coarse masks are created through dilation and edge disturbance, as illustrated in Figure 5. For stationary watermarks, we randomly choose a location within the video to composite them. For moving watermarks, a trajectory of movement that matches the length of the video clip is randomly created, and the watermarks are subsequently blended into the video frame-by-frame. For evaluation, we tested on 508 YouTube-VOS videos and 50 out of 90 DAVIS testing videos, using 100 colored watermarks: 40 selected from CLWD [10] and 60 randomly chosen from 160 in ILAW. The data synthesis process is consistent with training to ensure comparability. All videos are resized to 432 × 240 for both training and evaluation.

2) Training Details: We employ the Adam optimizer [43] to train all the model components. The temporal backbone trains with a batch size of 4 and an initial learning rate of 0.0001, running for 800k iterations. Additionally, we set the local sequence length to 10 and use horizontal flips for data augmentation. For the framework architectures, we use 4 blocks for the MQGA-Transformer and 8 MSVT [3] blocks for the temporal backbone.

3) Evaluation Metrics: Similar to [2] and [3], we employ peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [44], RMSE$_w$, and video-based Fréchet inception distance (VFID) [45] as evaluation metrics to assess the performance of visible video watermark removal approaches. RMSE$_w$ denotes the root-mean-square error metric applied to the watermarked region. PSNR, SSIM, and RMSE$_w$ are used to measure low-level similarity, while VFID is employed to gauge both high-level perceptual quality and temporal consistency between the output and the ground truth.

4) Comparison Methods: We compared DECO with 6 state-of-the-art (SOTA) video methods, including Shift-Net [2], RVRT [12], and EMVD [13] for video restoration, as well as FuseFormer [41], E^{2}FGVI [14], and ProPainter [3] for video inpainting. Meanwhile, we also compared DECO with the SOTA image watermark removal methods, including WDNet [10], DENet [5], SplitNet [4], and SLBR [1]. Re-implementations of all algorithms are conducted following their default experimental setups.

B. Comparisons With the State-of-the-Arts

1) Quantitative Evaluation: Table I summarizes the DECO framework's quantitative performance against existing SOTA methods on YouTube-VOS and DAVIS. Our approach notably surpasses others in key metrics for video restoration and inpainting. Specifically, DECO improves PSNR by 3.67 and 2.78, and SSIM by 0.0171 and 0.0143 on YouTube-VOS and DAVIS, respectively, compared to the Shift-Net baseline. On the VFID metric, DECO exceeds ProPainter and Shift-Net, reducing scores to 0.050 and 0.117 from 0.107 and 0.176 respectively, showcasing its superiority and generalizability in watermark removal while maintaining temporal consistency. Due to the large variation of watermark shape and appearance, video restoration methods like EMVD and RVRT which merely rely on the network's content recovery capability perform worse than video inpainting methods such as FuseFormer and E^{2}FGVI, which employ binary masks for specifying the location of watermarks. To demonstrate the necessity and superiority of the video watermark removal method, we process videos as sequences of watermarked images, applying established image watermark removal methods to each frame individually. We report the metric results of these methods in Table III. Although these methods can produce relatively high-quality background reconstruction images, they yield high VFID scores, indicating inferior temporal consistency due to the inability to exploit temporal information. Our method outperforms these methods that are specifically designed to cope with watermark removal on all evaluation metrics.

2) Qualitative Evaluation: Figure 7 presents our qualitative comparisons with existing video restoration and inpainting methods. The results of our method have superior visual quality in watermark removal and minimized artifact and blur issues. Existing video restoration methods, including RVRT and EMVD often fail to completely eliminate watermarks due to significant damage, whereas existing video inpainting methods including FuseFormer, E^{2}FGVI, and ProPainter, suffer from loss of important background content and fill the corrupted regions with other artifacts. In contrast, our DECO approach addresses these challenges, providing a balanced solution for watermark removal and video restoration. Shift-Net, while competitive produces unwanted colored noise in the second to sixth examples. Besides, Figure 6 presents a qualitative comparison between our method and three other competitive image watermark removal methods. We observed that their results exhibited unpleasant flickering and deficient content coherence. These approaches fail to address the challenges of moving watermarks and generate temporally consistent results caused by the inability to model temporal information. For instance, SplitNet performs well on the first and third frames but leaves watermark traces in the second frame.

3) User Study: We conduct a user study to compare the effectiveness of removing stationary and moving video watermarks. We select the best video restoration method, Shift-Net [2]), and the best video inpainting method, ProPainter [3], for comparison. Fifteen volunteers are invited to grade 30 videos randomly chosen from the YouTube-VOS and DAVIS test sets. Each test sample consists of a quadruple: the input

TABLE III
QUANTITATIVE COMPARISONS WITH SOTA IMAGE WATERMARK REMOVAL MODELS ON DAVIS DATASETS

Method	PSNR↑	SSIM↑	RMSE$_w$↓	VFID↓
WDNet [10]	18.63	0.7324	26.852	0.976
DENet [5]	27.68	0.9371	10.778	0.293
Splitnet [4]	28.00	0.9391	11.289	0.345
SLBR [1]	29.44	0.9481	9.655	0.253
DECO (Ours)	**33.34**	**0.9777**	**6.028**	**0.117**

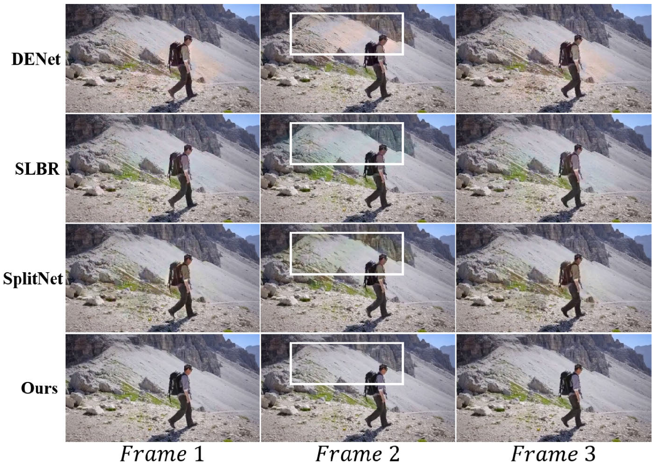

Fig. 6. Qualitative comparisons of our proposed DECO with other SOTA watermark removal methods. The inability of image models to model temporal information leads to temporally inconsistent results.

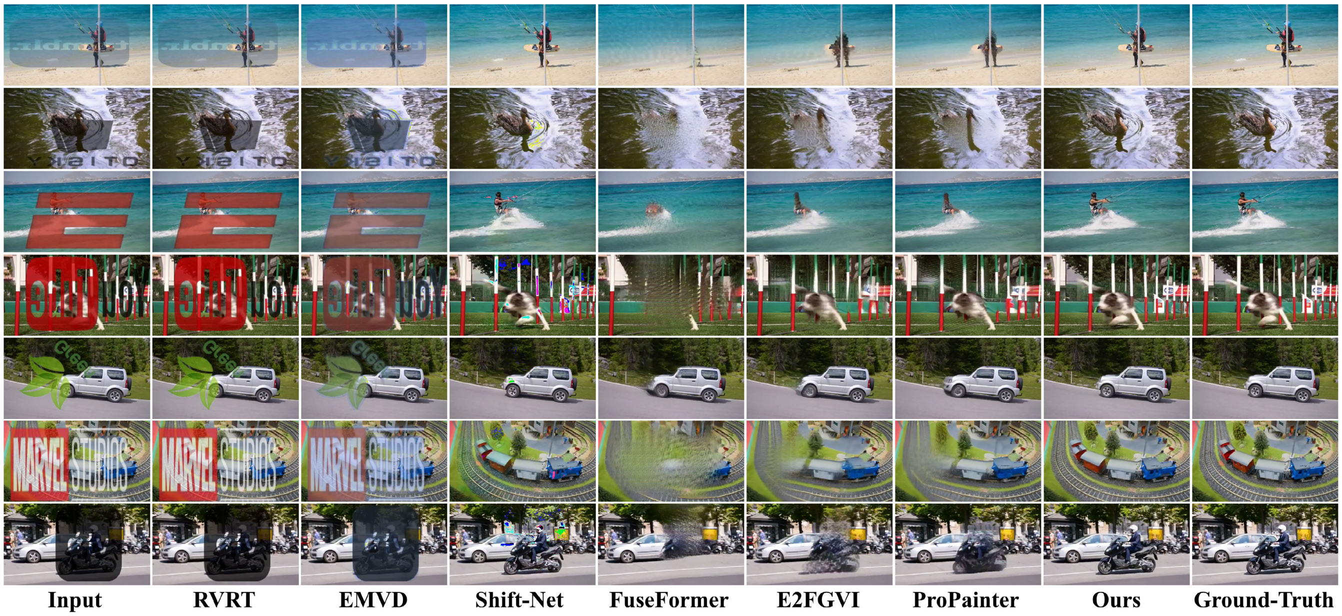

Fig. 7. Qualitative comparisons between our proposed DECO and other previous SOTA methods. Best viewed zoomed in.

TABLE IV
USER STUDY RESULTS

Method	Shift-Net	ProPainter	Ours
Rank 1	5	0	25
Rank 2	20	6	4
Rank 3	5	24	1

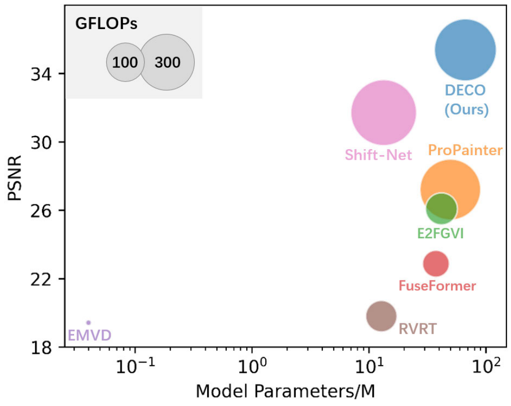

Fig. 8. Comparison of model parameters and GFLOPs of DECO and other SOTA methods. DECO outperforms other methods while maintaining efficiency.

watermarked video, the output of our method, and the outputs of two comparison methods, with the video orders randomly shuffled. Volunteers rank the results of the three methods based on the video's visual quality, considering criteria such as background content quality, temporal consistency, and presence of noise and artifacts. As shown in Table IV, we count the number of videos corresponding to each ranking for each method, which suggests that the volunteers prefer results of our method on majority of test samples.

4) Efficiency Comparison: Figure 8 shows the number of model parameters and GFLOPs for all video methods. The computationally intensive of our method is slightly less than Shift-Net but achieving better background recovery performance.

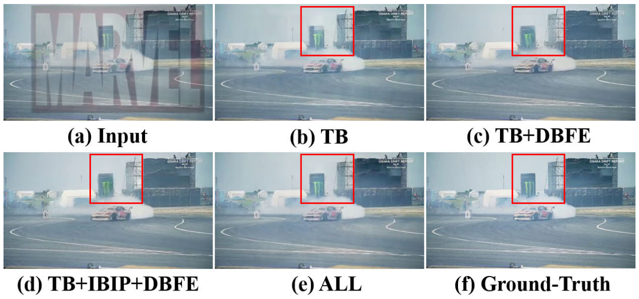

Fig. 9. Qualitative comparisons of our proposed DECO with different combination types of model components. "TB": the temporal backbone. "ALL": TB+IBIP+DBFE+FFSM. Zoom in for a better view.

C. Ablation Analysis

To provide deeper insights into the effectiveness of each component and the training strategy of the proposed DECO method, we conduct extensive ablation experiments on the DAVIS dataset. The results of these ablation experiments are summarized in Table V and VI, Figure 9 and 10.

1) Effect of IBIP and DBFE of the FPFE Module: To better understand the effectiveness of the FPFE module, we investigate its sub-modules, IBIP and DBFE, with findings detailed in Table V and Figure 9. Table V shows that using IBIP alone (row M-(2) vs. M-(1)) diminishes background recovery quality. The possible reason is that the predictions of IBIP disturbs the appropriate modelling of temporal dependencies. Conversely, integrating the backbone with frame-wise prior features from DBFE (row M-(3) vs. M-(1)) mitigates these challenges, enhancing performance on both PSNR and SSIM metrics. Additionally, removing watermark interference via

TABLE V

ABLATION STUDY RESULTS OF THE PROPOSED DECO WITH DIFFERENT COMBINATIONS OF COMPONENTS ON DAVIS DATASET

M-(#)	Backbone	IBIP	DBFE	FFSM	PSNR↑	SSIM↑
1	√	×	×	×	32.17	0.9748
2	√	√	×	×	31.67	0.9691
3	√	×	√	×	32.75	0.9757
4	√	√	√	×	32.93	0.9758
5	√	√	√	√	**33.34**	**0.9777**

TABLE VI

ABLATION STUDY RESULTS OF THE PROPOSED DECO WITH DIFFERENT TYPES OF TEMPORAL BACKBONE ON DAVIS DATASET

M-(#)	Backbone	PSNR↑	SSIM↑
1	FuseFormer [41]	32.82	0.9744
2	MSVT [3]	**33.34**	**0.9776**

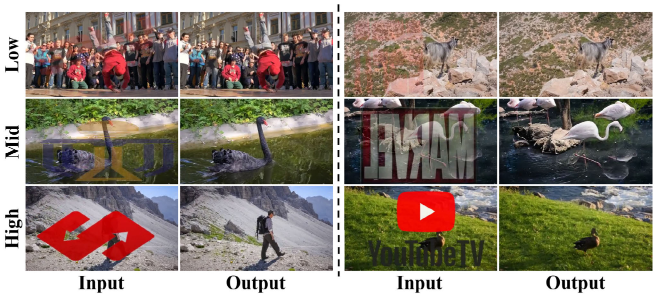

Fig. 10. Qualitative comparisons of our proposed DECO in handling video watermarks with different opacity levels. “Low”, “Mid” and “High” correspond to low, middle, and high levels of watermark opacity, respectively, with their associated opacity levels, α, ranging within [0.1, 0.25], [0.45, 0.65], and [0.9, 0.95].

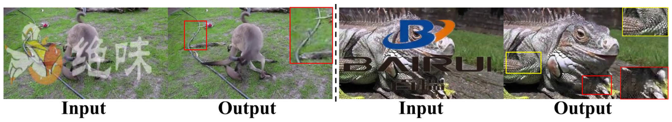

Fig. 11. Two failure cases. The key regions are zoomed in for better illustration.

IBIP (row M-(3) vs. M-(4)) and incorporating DBFE’s prior information (row M-(4) vs. M-(2)) significantly improves the performance, indicating that image-derived information can boost background recovery quality. Figure 9 provides a comprehensive qualitative comparison of these components. As indicated in the red box, the incorporation of the DBFE (c) helps improve the background restoration performance, and the adoption of IBIP (d) contributes to more accurate recovery results.

2) Effect of FFSM: Table V and Figure 9 showcase the importance of the FFSM in our DECO, highlighting both the quantitative and qualitative impacts of FFSM. Specifically, the comparison in Table V (row M-(5) vs. M-(4)) evidences a performance drop without FFSM, underscoring its critical role. Furthermore, in subfigure (e) of Figure 9, the addition of FFSM helps clarify the building edges. The reason is that FFSM can enhance the robustness across diverse watermarks. Additionally, Figure 10 shows DECO’s capability in handling watermarks of varying opacities, mainly owing to the adoption of FFSM.

3) Effect of the Temporal Backbone: We assess the generality of DECO by substituting the MSVT [3] backbone with FuseFormer [41]. Table VI reveals negligible performance drops after replacing MSVT with FuseFormer. This demonstrates DECO’s compatibility with various temporal backbones, and underscores the versatility and generality of DECO.

4) Effect of Training Strategy: To further comprehend the proposed DECO’s design and training strategies, we conduct more ablation studies with five variants of DECO. The results are presented in Table VII. The results suggest that the enhancement achieved by our proposed method stems not from an increase in model parameters, but from the specialized design of each module.

To validate the necessity of independently pretraining IBIP and DBFE (as outlined in Line 2 and Line 3 of Algorithm 1), we initially train the entire model, including IBIP and DBFE, from scratch in an end-to-end manner. The experimental result in row M-(1) of Table VII indicates a significant performance decline. This decline is probably due to the absence of targeted supervision for IBIP and DBFE during training, which hinders the effective extraction of prior features needed to assist the temporal backbone.

To further verify the necessity of separately pretraining IBIP and DBFE, we combine their pretraining processes by pretraining the FPFE module on both the CLWD and ILAW datasets end-to-end. The experimental results, shown in row M-(3) of Table VII, exhibit worse performance compared to row M-(6). This may be due to the joint pretraining strategy’s inability to make IBIP focus on erasing watermark patterns and predicting intrinsic background images, thereby adversely impacting the performance of the temporal backbone.

To emphasize the importance of pretraining IBIP and DBFE on the CLWD and ILAW datasets (as outlined in Line 2 and Line 3 of Algorithm 1), we conduct an experiment where IBIP and DBFE are pretrained directly using frame images from the video datasets. The result in row M-(2) of Table VII indicates inferior performance compared to M-(6). This decline is attributed to the limited scene diversity in video datasets, which restricts the ability of IBIP and DBFE to learn sufficient prior knowledge, while the CLWD and ILAW datasets have more diverse scenes.

To validate the rationale behind the dual-branch design of the KR and ST branches, we conducted experiments by individually removing each branch from DBFE. The results, shown in rows M-(4) and M-(5) of Table VII, indicate certain degrees of performance reduction after removing either branch. Notably, the removal of the KR branch results in a more significant decrease, suggesting that the KR branch plays a more critical role in frame restoration.

D. Limitation and Discussion

Failure Cases: In Figure 11, we show two failure cases of restorations. The left side of Figure 11 shows that when watermarks feature nested patterns with varying opacities,

TABLE VII

Ablation Study Results of Training Strategy of DECO. "Pretrain." Represents the Pretraining of IBIP and DBFE. "Img. Dataset" Represents the CLWD and ILAW Datasets for the Pretraining of IBIP and DBFE. "Indep." Represents Independently Pretraining IBIP and DBFE of FPFE Module. "KR" and "ST" Represent the KR and ST Branches of DBFE in FPFE, Respectively

M-(#)	Pretrain.	Img. Dataset	Indep.	KR	ST	PSNR↑	SSIM↑
1	×	×	×	√	√	26.90	0.9117
2	√	×	√	√	√	32.15	0.9658
3	√	√	×	√	√	28.02	0.9167
4	√	√	√	×	√	32.43	0.9632
5	√	√	√	√	×	32.64	0.9663
6	√	√	√	√	√	**33.34**	**0.9777**

particularly in the presence of moving watermarks, the restoration performance deteriorates. Additionally, when the background texture is intricate, the model may struggle to restore high-frequency details, leading to blurring and artifacts, as shown on the right side of Figure 11. This shows that these situations are still challenging for video watermark removal. Furthermore, the experimental results indicate that watermarks with the following characteristics are more resistant to attacks: complex patterns, parts with varying opacities, and movement, which offers insights for the design and application of reliable watermarks.

V. Conclusion

In this paper, we propose a novel framework named DECO to address the challenge of visible video watermark removal. DECO effectively extracts frame-wise prior features by leveraging common-sense knowledge and residual background information, while accurately modeling the temporal dependencies necessary for video content restoration. Extensive experiments and comprehensive ablation studies demonstrated the superiority and generality of DECO. Our method not only overcomes the limitations of existing video restoration techniques in handling complex watermark noise but also significantly improves the quality of the recovered image content. This innovation and advancement have the potential to evaluate the robustness and security of watermarks, offering important practical application value. Nonetheless, our study has certain limitations. For example, watermarks consisting of nested patterns with varying opacities have not been addressed. These limitations indicate areas for improvement and further investigation, which we plan to pursue in the future.

References

[1] J. Liang, L. Niu, F. Guo, T. Long, and L. Zhang, "Visible watermark removal via self-calibrated localization and background refinement," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4426–4434.

[2] D. Li et al., "A simple baseline for video restoration with grouped spatial–temporal shift," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9822–9832.

[3] S. Zhou, C. Li, K. C. K. Chan, and C. Change, "ProPainter: Improving propagation and transformer for video inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10477–10486.

[4] X. Cun and C.-M. Pun, "Split then refine: Stacked attention-guided resunets for blind single image visible watermark removal," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1184–1192.

[5] R. Sun, Y. Su, and Q. Wu, "DENet: Disentangled embedding network for visible watermark removal," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 2, pp. 2411–2419.

[6] J. Xu et al., "Video dehazing via a multi-range temporal alignment network with physical prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18053–18062.

[7] H. Chen et al., "Snow removal in video: A new dataset and a novel method," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13165–13176.

[8] J. Wang, W. Weng, Y. Zhang, and Z. Xiong, "Unsupervised video deraining with an event camera," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10797–10806.

[9] L. Niu, X. Zhao, B. Zhang, and L. Zhang, "Fine-grained visible watermark removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12724–12733.

[10] Y. Liu, Z. Zhu, and X. Bai, "WDNet: Watermark-decomposition network for visible watermark removal," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3684–3692.

[11] D. Huo, Z. Zhang, H. Su, G. Li, C. Fang, and Q. Wu, "WMFormer++: Nested transformer for visible watermark removal via implict joint learning," 2023, *arXiv:2308.10195*.

[12] J. Liang et al., "Recurrent video restoration transformer with guided deformable attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 378–393.

[13] M. Maggioni, Y. Huang, C. Li, S. Xiao, Z. Fu, and F. Song, "Efficient multi-stage video denoising with recurrent spatio-temporal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3465–3474.

[14] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17562–17571.

[15] J. Liang et al., "VRT: A video restoration transformer," *IEEE Trans. Image Process.*, vol. 33, pp. 2171–2182, 2024.

[16] J. Wu, X. Yu, D. Liu, M. Chandraker, and Z. Wang, "DAVID: Dual-attentional video deblurring," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2365–2374.

[17] F. Li, H. Bai, and Y. Zhao, "Learning a deep dual attention network for video super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 4474–4488, 2020.

[18] Z.-S. Liu, L.-W. Wang, C.-T. Li, W.-C. Siu, and Y.-L. Chan, "Image super-resolution via attention based back projection networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3517–3525.

[19] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3360–3369.

[20] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[21] N. Xu et al., "YouTube-VOS: Sequence-to-sequence video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 585–601.

[22] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 724–732.

[23] J. Ren, Q. Zheng, Y. Zhao, X. Xu, and C. Li, "DLFormer: Discrete latent transformer for video inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3511–3520.

[24] 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.

[25] E. Lee, J. Yoo, Y. Yang, S. Baik, and T. H. Kim, "Semantic-aware dynamic parameter for video inpainting transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12903–12912.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 234–241.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] Y. Chang and C. Jung, "Single image reflection removal using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1954–1966, Apr. 2019.

[29] C. Chen et al., "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1329–1338.

[30] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 6309–6318.

[31] C. Zheng and A. Vedaldi, "Online clustered codebook," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22741–22750.

[32] K. Liu, Y. Jiang, I. Choi, and J. Gu, "Learning image-adaptive codebooks for class-agnostic image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5350–5360.

[33] Q. Liu et al., "Reduce information loss in transformers for pluralistic image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11337–11347.

[34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[36] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[40] 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.

[41] R. Liu et al., "FuseFormer: Fusing fine-grained information in transformers for video inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Sep. 2021, pp. 14040–14049.

[42] 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. 2017.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diega, CA, USA, 2015, pp. 1–11.

[44] W. Zhou, 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.

[45] T.-C. Wang et al., "Video-to-video synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Aug. 2018, pp. 1144–1156.

[46] Y. Leng, C. Fang, G. Li, Y. Fang, and G. Li, "Removing interference and recovering content imaginatively for visible watermark removal," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 4, pp. 2983–2990.

Junye Chen received the B.Eng. degree from the South China University of Technology in 2023. He is currently pursuing the Ph.D. degree with Sun Yat-sen University. His research interests include computer vision and image processing.

Chaowei Fang (Member, IEEE) received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2019. He is currently an Associate Professor with the School of Artificial Intelligence, Xidian University, Xi'an, China. He has contributed as an author or coauthor to more than 40 publications published in prestigious journals and conferences. His research interests include low-level image processing, medical image analysis, and machine learning. He served as a Senior Program Committee Member for ECAI 2024.

Jichang Li received the M.Eng. degree from the School of Computer Science and Technology, South China University of Technology, in 2020, and the Ph.D. degree in computer science from The University of Hong Kong, in 2024. He is currently an Assistant Researcher at the Pengcheng National Laboratory, Shenzhen, China. His research interests focus on computer vision and deep learning. Additionally, he serves as a reviewer for numerous academic journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, CVPR, ICCV, and ECCV.

Yicheng Leng received the bachelor's and master's degrees in science from Chinese University of Hong Kong, Shenzhen, in 2021 and 2023, respectively. His research interests include deep learning application in computer vision and medical imaging.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a Full Professor with the School of Computer Science and Engineering, Sun Yat-sen University. He has authored or co-authored more than 150 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of the ICCV 2019 Best Paper Nomination Award. He serves as the Area Chair for CVPR 2024. He has been serving as a reviewer for numerous academic journals and conferences, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, *International Journal of Computer Vision*, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Cybernetics, CVPR, ICCV, ECCV, and NeurIPS.