CIRI: Curricular Inactivation for Residue-aware One-shot Video Inpainting

Weiying Zheng [1*] Cheng Xu [1*] Xuemiao Xu[1,2,3,4†] Wenxi Liu [5] Shengfeng He [6†]
[1]South China University of Technology [2]State Key Laboratory of Subtropical Building Science
[3]Ministry of Education Key Laboratory of Big Data and Intelligent Robot
[4]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[5]Fuzhou University [6]Singapore Management University

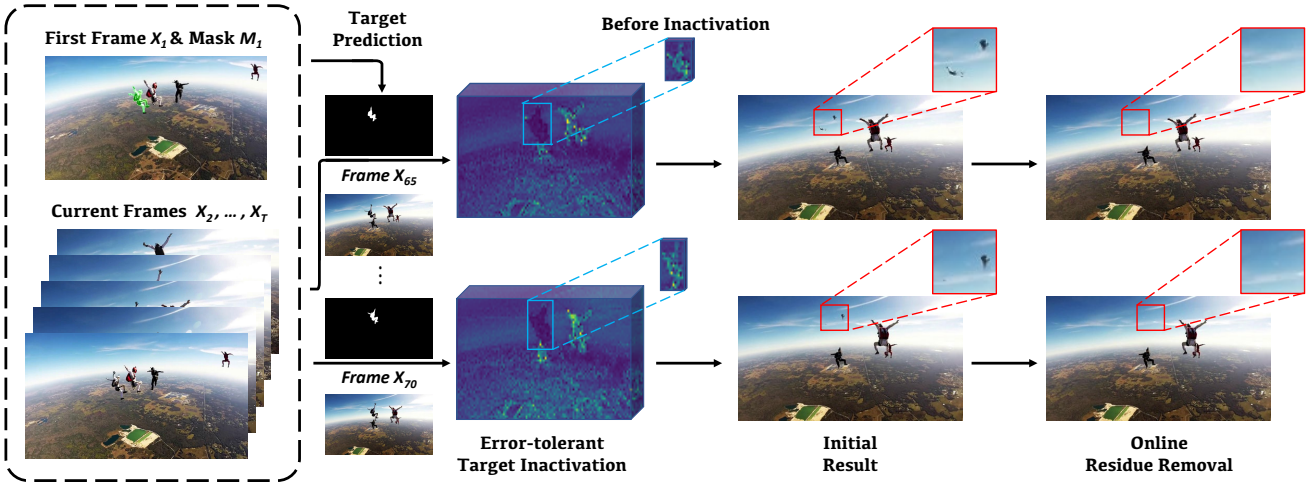

Figure 1: We resolve a challenging problem of video inpainting with the one-and-only target annotation in the first frame. Our method is completely plug-and-play that allows converting existing inpainting models to the one-shot setting with significant performance improvements.

Abstract

Video inpainting aims at filling in missing regions of a video. However, when dealing with dynamic scenes with camera or object movements, annotating the inpainting target becomes laborious and impractical. In this paper, we resolve the one-shot video inpainting problem in which only one annotated first frame is provided. A naive solution is to propagate the initial target to the other frames with techniques like object tracking. In this context, the main obstacles are the unreliable propagation and the partially inpainted artifacts due to the inaccurate mask. For the former problem, we propose curricular inactivation to replace the hard masking mechanism for indicating the inpainting target, which is robust to erroneous predictions in long-term video inpainting. For the latter, we explore the properties of inpainting residue and present an online residue removal method in an iterative detect-and-refine manner. Extensive experiments on several real-world datasets demonstrate the quantitative and qualitative superiorities of our proposed method in one-shot video inpainting. More importantly, our method is extremely flexible that can be integrated with arbitrary traditional inpainting models, activating them to perform the reliable one-shot video inpainting task. Video demonstrations can be found in our supplement, and our code can be found at https://github.com/Arise-zwy/CIRI.

1. Introduction

Video inpainting aims at filling holes with plausible content that is spatially and temporally consistent with the original video. It serves as an essential and fundamental function in numerous video editing applications, such as scratch restoration [14, 36], video retargeting [1, 20], and object removal [18].

Existing methods have a strong assumption that the in-

*Both authors contributed equally to this research.
†Corresponding authors: Xuemiao Xu (xuemx@scut.edu.cn) and Shengfeng He (shengfenghe@smu.edu.sg).

painting targets are well-defined across all the frames, and thus their research focus lies on connecting individual inpainting results with coherency [16, 19, 21, 22, 44]. However, this problem definition is too ideal that cannot be realized in practical scenarios. A short video with a few seconds can contain hundreds of frames, and annotating all of them for every editing task is obviously laborious and infeasible.

In this paper, we aim to resolve a challenging video inpainting problem that only the annotation of the first frame is available. This practical setting has not been studied in more depth, but it is explored from an object segmentation perspective. For instance, existing one-shot inpainting models [18, 28] focus only on improving the mask propagation quality. However, they neglect the fact that the predicted/propagated masks cannot be always perfect.

Here we address this dilemma from a pure inpainting aspect. We embrace the truth that predicted masks are erroneous, and therefore concentrate on coping with these inaccuracies in the inpainting framework. To this end, we propose a **C**urricular **I**nactivation framework for **R**esidue-aware one-shot video **I**npainting (**CIRI**). First, we observe that the main source of inpainting errors comes from the hard masking mechanism. It assumes all target pixels are included in the mask and no others are left outside, which is obviously false in the one-shot setting. We therefore, tailor a new target indication mechanism, curricular inactivation, to tolerate inaccurate target masks. Specifically, instead of masking out the target regions in the image space that might ruin good image content, we propose to inactivate multi-scale feature responses of the predicted areas. More importantly, we introduce a dual-curriculum learning strategy into our inactivation to gradually transfer the dependency from the perfect ground truth to the inaccurate mask. Second, it is inevitable to have partial regions unsuccessfully inpainted when using an incomplete mask, and these regions are shown in different data distributions to either foreground or background. We therefore design an online residue removal scheme to actively detect these artifacts and remove them iteratively during the inference stage. Fig. 1 shows our overall pipeline of one-shot video inpainting.

To quantitatively evaluate the proposed method, we construct a synthetic object removal dataset based on Youtube-VOS [41], DAVIS [29], and OVID [30]. Extensive experiments on this synthetic dataset as well as other real-world datasets demonstrate superior performance over state-of-the-art traditional and one-shot video inpainting methods. Another important feature of our method is that the proposed two main modules are completely plug-and-play that can convert traditional methods to one-shot inpainting and significantly boost their performances.

In summary, our main contributions are threefold:

- We propose a curricular inactivation strategy to substitute hard masking for indicating inpainting targets in an unreliable input scenario. On one hand, it inactivates multi-scale feature responses to prevent destroying original good image content. On the other hand, we use curriculum learning to progressively tolerate imperfect input masks.
- We explore the property of inpainting artifacts, and present an online learning strategy for iteratively detecting and removing inpainting residues in the inference phase.
- The proposed method not only demonstrates superior performances over state-of-the-art inpainting methods, but also enables converting arbitrary inpainting models to the one-shot setting with a significant performance improvement.

2. Related Work

Traditional Video Inpainting. Traditional video inpainting seeks to fill the corrupted regions in each frame while maintaining spatial-temporal consistency across frames. Early works tackle this problem via patch matching between frames [9, 11]. Recently, deep learning-based methods have dominated this field. They mainly focus on exploiting the spatial-temporal cues from neighboring frames by either using dense correspondences [10, 16, 19, 21, 42, 40], attentions [22, 23, 27, 44], or 3D convolutions [5, 47]. Several methods [28, 31, 45] also resort to internal learning for propagating information of known regions to unknown parts in a single video. However, traditional video inpainting methods assume ground truth target masks of each frame are accessible during the testing, and may suffer from significant performance degradation once the provided target masks are inaccurate. This largely limits their real-world applications.

Low-shot Video Inpainting. To sidestep the need for ground truth masks, low-shot video inpainting endeavors to inpaint the target regions with the annotations of only a few frames. Trinh *et al*. [18] first predicts target masks with a few user strokes as guidance, and then propagates these masks to all remaining frames, allowing for automatic video inpainting. IIVI [28] adopts an internal learning scheme to achieve mask propagations with a single frame mask. Wu *et al*. [38] jointly learn a mask prediction network and another completion network by applying a cycle-consistency regularization. Although ground truth masks are no more required, the above methods tend to yield inferior results due to inaccurate mask predictions or propagations. In contrast, we make the first attempt to introduce the spirit of curriculum learning to solve the one-shot video inpainting problem. We propose a novel curricular target inactivation mechanism to achieve considerable tolerance to inaccurate target masks, leading to more faithful and accurate inpainting results.

Low-shot Video Object Segmentation. Predicting accurate target masks of each frame with limited annotations

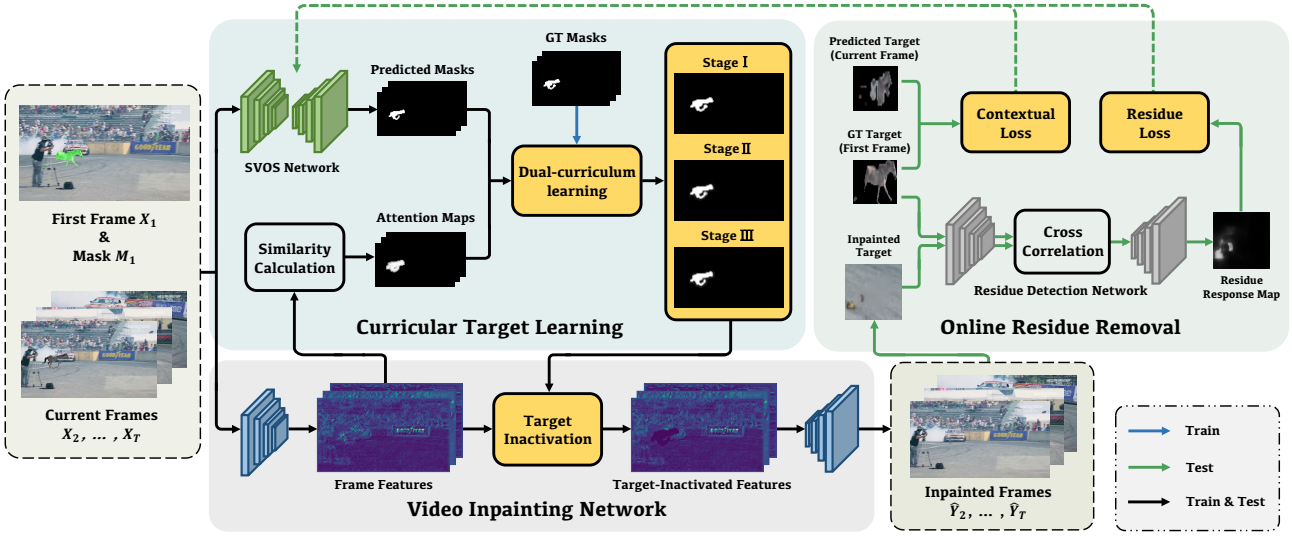

Figure 2: Our framework consists of two modules: the curricular target inactivation module for error-tolerant inpainting with inaccurate masks, and the online residue removal module to suppress inpainting artifacts during testing. Notably, during testing, only the SVOS network is trainable, while both the video inpainting network and residue detection network are frozen. Conversely, during training, only the video inpainting network is trainable.

is a critical step to realizing low-shot video inpainting. Existing methods mainly explore spatial-temporal relations to assist the segmentation of different frames. For example, several works [3, 4, 15, 35] achieve this goal by fine-tuning a pretrained model with limited image-mask paired data from target domains. There are also other attempts to use feature matching [7, 13, 34] and memory mechanisms [8, 26] to facilitate the segmentation across frames. Various methods [6, 25, 43] borrow the information from neighboring frames by utilizing the optical flow, which effectively boosts the low-shot segmentation performance. Differently, we propose an online residue removal scheme by iteratively fine-tuning the pretrained semi-supervised video object segmentation (SVOS) model with the awareness of minimizing the residues, which effectively improves segmentation performance and suppresses the inpainting artifacts.

3. Method

3.1. Overview

Problem Formulation. Given a video sequence $X := \{X_t \in \mathbb{R}^{H\times W\times 3} | t = 1, 2, ..., T\}$ with length T, and a mask $M_1 \in \mathbb{R}^{H\times W\times 1}$ indicating the target region in the first frame, our goal is to synthesize the inpainted video $\widehat{Y} := \left\{\widehat{Y}_t \in \mathbb{R}^{H\times W\times 3} | t = 1, 2, ..., T\right\}$ with faithful content in the target region with spatial and temporal consistency across all frames, by using the one-and-only target annotation in the first frame. For training, the model requires the corrupted video sequences X, the ground truth masks of each frame M, and the ground truth inpainted video Y.

Network Design. The pipeline of our CIRI framework is shown in Fig. 2. Built upon an arbitrary traditional video inpainting backbone, our framework consists of two modules, including a curricular target inactivation module and a residue removal module, which cope with target feature inactivation and residue artifacts removal, respectively. Taking X and M_1 as input, the encoder of the inpainting network first extracts the features of each frame from X. These features are then sent to the curricular inactivation module for selectively inactivating multi-scale feature responses of the targets, by combining the reciprocal strengths of ground truth masks and learned masks. Afterwards, the target-inactivated features are fed to the decoder to render a faithful inpainted video. During the testing phase, the residue removal module actively detects the inpainting artifacts in the inpainted video and removes them iteratively. We discuss each component of our method in the following sections.

3.2. Curricular Inactivation for Video Inpainting

Target Inactivation. Earlier works [18, 28] mainly adopt a hard masking mechanism for video inpainting. They assume that all target pixels are included in the mask and no others are left outside. However, as the predicted masks cannot be always perfect, this assumption obviously does not hold in the one-shot setting, leading to unsatisfactory inpainting results. Thus, instead of directly applying a hard mask to mask out the target in the image space that may ruin the image content, we propose a target inactivation mechanism by selectively inactivating the multi-scale feature responses of the predicted target regions. The rationale behind this is that high-level features capture long-range and multi-scale

spatial correlations among the pixels of the input image due to their large receptive fields. For this reason, substantial information of the target region can remain in the pixels outside the target mask. This feature-level multi-scale masking strategy allows the model to fully exploit the meaningful target information to assist error-tolerant inpainting.

As shown in the bottom part of Fig. 2, given the input frame features F_{input} produced by the encoder of the inpainting network and an inactivation map M_{iac} with pixel values ranging from $[0, 1]$, we can obtain the target-inactivated output features F_{out} by operating as follows:

$$F_{out} = (1 - M_{iac}) \odot F_{input}, \tag{1}$$

where $\odot$ denotes the element-wise multiplication operation. By selectively inactivating the target feature responses, F_{out} not only well preserves the background features, but also retains meaningful responses that are beneficial for the subsequent target inpainting.

Dual-Curriculum Learning. To precisely inactivate the target responses, an ideal way is to use the ground truth masks of each frame as the inactivation maps. However, this may result in the overdependence of the model on ground truth masks. To alleviate this problem and boost the learning performance on inaccurate masks, we propose to equip the target inactivation mechanism with a dual-curriculum learning scheme (Fig. 3). This scheme shares a similar spirit to curriculum learning [2, 32, 46, 12], which allows the model to start with learning simple tasks and then gradually transfer to difficult tasks for better optimization.

In particular, our dual-curriculum learning scheme comprises two curriculum tasks by initially learning from the ground truth masks and then gradually transferring the learning focus to the inaccurate masks. To be more specific, each curriculum task includes three training stages. In the first stage, the model is trained with the inactivation maps learned from the ground truth masks only, which offers precise guidance to grant the model an initial capability of faithful target inpainting. However, over-reliance on ground truth masks can lead to significant performance degradation during the inference phase, where the ground truth masks are not available. We thus gradually replace the ground truth masks with inaccurate masks by interpolating the two for training in the second stage. By this means, our model can gain increasing robustness to the perturbations from the inaccurate masks. To fully evade the need for ground truth masks and enable the model to adapt to inaccurate masks, we only utilize the inaccurate masks for the third stage of training. The curriculum learning process can be formulated as follows:

$$M_c = \begin{cases} M_{gt} & 0 < e \leq n_1 \\ \alpha \odot M_{gt} + (1 - \alpha) \odot M_{err} & n_1 < e \leq n_2 \\ M_{err} & e > n_2, \end{cases} \tag{2}$$

where M_{gt}, M_{err}, and M_c denote the ground truth mask, inaccurate mask, and candidate inactivation mask, respectively. α is a curriculum learning rate decreasing from 1 to 0 for controlling the tradeoff between the ground truth and the inaccurate masks during training. e is the training epoch number of each stage controlled by two fixed parameters n_1 and n_2. Note that we only use the inaccurate mask as the candidate inactivation map during testing.

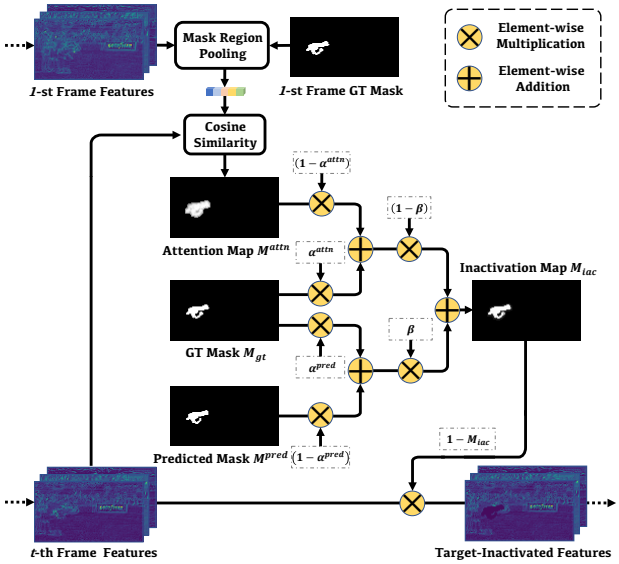

Figure 3: Illustration of our curricular target inactivation module. It gradually shifts learning focus from ground truth masks to predicted masks and complementary attention maps for better tolerance to inaccurate masks.

To obtain more reasonable inactivation maps, we introduce two kinds of inaccurate masks to inject rich and diverse target mask information for learning. The first is the mask M^{pred} predicted by a pretrained SVOS network on a frame, it provides an initial target mask which may contain undesired prediction errors. To provide complements and rectifications to the erroneously predicted masks, we further introduce the complementary attention map M^{attn} that indicates the possible target pixels, as the second source of the inaccurate masks. Specifically, to obtain the complementary attention map of a current frame, we first perform a mask region pooling on the first frame X_1 with its corresponding ground truth mask M_1 to get the feature vector of the target region. Then we compute the cosine similarity between the target feature vector and the feature vector in each feature pixel of the current frame. The resulting similarity map exhibit higher responses in the regions that share higher feature similarity with the target of the first frame. This can be thus used as a reasonable indicator of possible target regions. To further suppress the distractions from the background while capturing the potential target regions to the greatest extent, we perform a dilation operation on M^{pred} and use the dilated M^{pred} to mask out the background pixels in the similarity map, yielding the final complementary attention map.

The predicted mask and complementary attention map are respectively sent to two different curriculum branches for learning. The candidate inactivation masks of the two branches are finally combined via a learnable parameter β to deliver the final inactivation map M_{iac} as follows:

$$M_{iac} = \beta \odot M_c^{pred} + (1 - \beta) \odot M_c^{attn}, \tag{3}$$

where M_c^{pred} and M_c^{attn} represent the candidate inactivation masks from the predicted mask and complementary attention map curriculum branches, respectively. By adaptively fusing target regions from the above two complementary kinds of inaccurate masks, the final inactivation map M_{iac} provides more comprehensive guidance for the subsequent target inactivation. Thanks to the dual-curriculum learning scheme, our model gains considerable tolerance to inaccurate masks, and this is the key to achieving reliable one-shot inpainting.

Losses for Curricular-Inactivation Inpainting. The curricular target inactivation module can be integrated into arbitrary video inpainting backbones and optimized simultaneously by the supervision signals from the backbones. Taking the FuseFormer [22] as an example, two losses are applied on the inpainted videos for optimizing the entire curricular-inactivation inpainting framework. Specifically, a reconstruction loss $\mathcal{L}_{rec}$ is used to ensure the content of the inpainted video $\widehat{Y}$ are consistent with the ground truth video Y, which can be formulated as a L1 loss:

$$\mathcal{L}_{rec} = \left\| Y - \widehat{Y} \right\|_1. \tag{4}$$

To guarantee the visual realism and temporal consistency of the inpainted video, a temporal PatchGAN discriminator [21] is adopted to provide the adversarial supervision:

$$\begin{aligned} \mathcal{L}_{adv} = & \mathbb{E}_{x \sim P_Y} \left[\log D(x) \right] \\ & + \mathbb{E}_{z \sim P_{\widehat{Y}}} \left[\log(1 - D(z)) \right], \end{aligned} \tag{5}$$

where P_Y and $P_{\widehat{Y}}$ denote the distributions of real and inpainted videos, respectively.

The total loss for optimizing the network is as follows:

$$\mathcal{L}_{inpainting} = \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{adv} \cdot \mathcal{L}_{adv}, \tag{6}$$

where λ_{rec} and λ_{adv} are weighting parameters to balance the two losses, which are set to 1 and 0.01, respectively.

3.3. Online Residue Removal

Although the curricular inactivation mechanism enables considerable tolerance to inaccurate masks and greatly improves the inpainting performance, there may still exist partial pixels unsuccessfully inpainted due to the imperfect prediction of the inactivation map. To alleviate such inpainted residues and elevate inpainting performance, we present an online residue removal scheme to actively detect and remove them iteratively during the testing phase.

Residue Detection. To remove the residues, we propose to first detect them by modifying an offline-trained visual object tracking model SiamMask [37] as the residue detection network, which compares a template image and a (larger) search image to yield a dense response map. In particular, considering the residues mainly reside inside/around the target regions, we perform a $w \times h$ crop centered on the target object of the first frame X_1 and a larger $(2w \times 2h)$ crop centered on the target region of the t-th inpainted frame $\widehat{Y}_t$ to avoid the interference from the background. Here, w and h denote the width and height of the minimum bounding rectangle of the target object/region. The two cropped patches are then resized to 127×127 and 255×255 respectively, which are later sent to the same feature extractor f_θ, yielding two feature maps. Finally, the cross-correlation map between the two feature maps is computed as follows:

$$M_{corr} = f_\theta(X_1^{crop}) * f_\theta(\widehat{Y}_t^{crop}), \tag{7}$$

where X_1^{crop} and $\widehat{Y}_t^{crop}$ are the cropped target regions of X_1 and $\widehat{Y}_t$, respectively. $*$ denotes the cross-correlation operator. M_{corr} is the obtained cross-correlation map, which holds the similarities between the target features in the first frame $f_\theta(X_1^{crop})$ with each spatial element of the target features in the t-th inpainted frame $f_\theta(\widehat{Y}_t^{crop})$. The cross-correlation map is then sent to the decoder to deliver the final residue response map M_{res}. The residue response map exhibits higher responses in the residue artifacts pixels, which is a good guide for the subsequent residue removal.

Residue Removal. With the residues detected, we then propose two simple yet effective losses to fine-tune the SVOS network for encouraging more accurate predicted target masks, which can effectively suppress the undesired residues. Specifically, considering that the target in the first frame shares high semantic similarity with the targets in the remaining frames, we apply a contextual loss [24] between the cropped target object in the first frame X_1^{crop} and the cropped predicted target region of the t-th frame X_t^{pred}. The contextual loss aims to maximize the feature similarity between the targets from two frames in terms of the local semantics and global context, which is computed as follows:

$$\mathcal{L}_{cx} = -\sum_{t=2}^{T} (\log CX(\Phi(X_1^{crop}), \Phi(X_t^{pred}))), \tag{8}$$

where CX represents cosine similarity computation between the context features $\Phi(\cdot)$ extracted from the "$conv3\_2$" layer of a pretrained VGG-19 [33]. By encouraging higher feature similarity between the groud truth target in the first frame and the predicted targets in the current frames, the contextual loss urges the SVOS network to predict more accurate target masks, thus relieving the inpainting residues.

Moreover, we also propose a residue loss to explicitly alleviate the inpainting residues by minimizing the sum of

Figure 4: Qualitative comparison on the YouTube-VOS (the first four rows) and DAVIS (the last two rows) datasets. (c)-(f) are the results of compared methods using predicted masks (b) generated by STCN [8], while (h) are the results of the one-shot video inpainting method using masks (the insets in (h)) produced by itself. Note that all competitors ((c)-(f)) are fed with masks detected by the same SVOS model as ours except for the first frame. Best viewed with zooming in digital version.

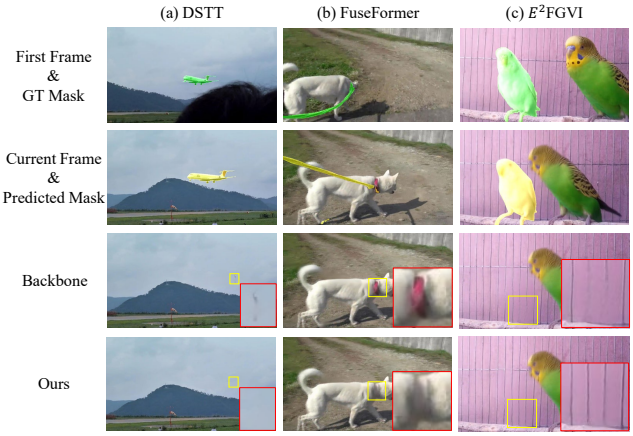

Figure 5: Qualitative comparison of different traditional backbones integrated with our framework on YouTube-VOS. Our framework effectively grants the traditional models considerable robustness to inaccurate target masks, leading to error-tolerant and convincing inpainting.

all the pixel values in the residue response map:

$$\mathcal{L}_{res} = \sum_{i=1}^{W}\sum_{j=1}^{H} M_{res}(i,j). \tag{9}$$

With the residue loss, the SVOS network produces more complete masks to ensure fewer target pixels remain outside the predicted masks. This further suppresses the residues.

In summary, the total loss for the online residue removal is a weighted sum of two losses:

$$\mathcal{L}_{removal} = \lambda_{cx} \cdot \mathcal{L}_{cx} + \lambda_{res} \cdot \mathcal{L}_{res}, \tag{10}$$

where λ_{cx} and λ_{res} are weighting parameters for loss terms, which are set to 0.1 and 1, respectively.

4. Experiments

4.1. Settings

Implementation details. To train the curricular inactivation framework, we choose a frozen pretrained STCN [8] as our SVOS network for target prediction. Following previous works [21, 22, 44, 39], we first pretrain the inpainting backbone with random masks on the training set of YouTube-VOS [41] for 150 epochs, ensuring initial inpainting capability. Next, we plug the curricular inactivation framework into the initialized backbone and train the entire network with our synthetic data for another 30 epochs. To mitigate the learning difficulty and achieve better convergence, n_1 of both curriculum branches are empirically set to 5, n_2 to 15 and 20 for the predicted mask branch and the complementary attention map branch, respectively. The learning rate is initialized as 2e-5 and decayed by 0.1 every 20 epochs.

During the online residue removal phase, we use a modified video object tracking model SiamMask [37] as the residue detection network and train it with our synthetic residue video sequences for 80k iterations. Then we freeze

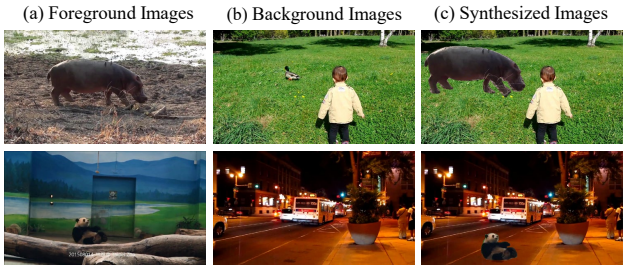

Figure 6: Examples of foreground/background images and their corresponding synthesized images for data synthesis.

Table 1: Quantitative comparison on the synthetic dataset.

Methods	Local			Global		
	PSNR↑	SSIM↑	VFID↓	PSNR↑	SSIM↑	VFID↓
VINet [16]	21.30	0.600	0.966	29.25	0.958	0.221
CAP [19]	24.81	0.749	0.917	31.21	0.972	0.172
LGTSM [5]	26.79	0.838	0.820	32.06	0.977	0.137
STTN [44]	26.02	0.791	0.839	35.13	0.986	0.120
IIVI* [28]	27.82	0.843	0.743	34.31	0.986	0.125
IIVI [28]	19.26	0.541	1.168	28.65	0.941	0.455
DSTT [23]	26.34	0.811	0.863	35.37	0.987	0.125
DSTT+Ours	27.48	0.851	0.826	36.11	0.989	0.116
FuseFormer [22]	27.21	0.842	0.838	35.93	0.989	0.107
FuseFormer+Ours	28.26	0.870	0.777	36.52	0.990	0.096
E^{2}FGVI [21]	28.11	0.865	0.764	36.45	0.990	0.089
E^{2}FGVI+Ours	**29.01**	**0.888**	**0.719**	**36.84**	**0.991**	**0.084**

* IIVI directly uses the masks predicted by STCN [8] as target guidance.

the inpainting network and the residue detection network, and fine-tune the SVOS network for 2 epochs on the input testing video with a learning rate of 1e-6. All the above networks are optimized by the Adam optimizer [17].

Datasets. Since there is no publicly available dataset for our setting, we synthesize the training/testing data from the existing datasets for video object segmentation, *i.e.*, YouTube-VOS [41], OVID [30], and DAVIS [29]. YouTube-VOS/OVID contains 3417/604, 474/140, and 508/154 videos used for training, validation, and testing, respectively. DAVIS is composed of 150 high-quality videos. Specifically, our training set begins by filtering out videos in YouTube-VOS where the annotated objects are extremely small or have brief appearances. We then randomly resize targets that occupy more than half of the image size and insert them into randomly chosen background video sequences. This process allows us to synthesize a total of 3417 training videos for experimentation. As for testing, we randomly select 90 foreground videos from OVID and paste them into 90 different background videos from DAVIS. Finally, we have 90 synthetic videos for quantitative evaluations. We also use 508 and 30 videos from the testing sets of YouTube-VOS and DAVIS respectively for qualitative evaluations. In Fig. 6, we showcase several examples of foreground/background images from the YouTube-VOS dataset and their corresponding synthesized images.

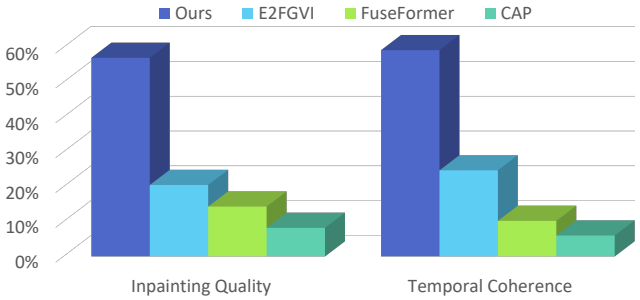

Figure 7: User study results. Our method significantly outperforms the other methods by producing more faithful and temporally consistent results.

Metrics. For quantitative evaluations, we consider three metrics including structure similarity measure (SSIM), peak signal-to-noise ratio (PSNR), and video-based Frechet inception distance (VFID). SSIM and PSNR are used for assessing the quality of overall reconstruction. VFID measures the spatial-temporal consistency and perceptual quality of the inpainting results. Additionally, we also measure the mean of region similarity and the contour accuracy $\mathcal{J}\&\mathcal{F}$ between the predicted masks and the corresponding ground truths to evaluate the mask prediction quality of the SVOS network in the ablation study.

4.2. Comparison with Existing Methods

Since very few existing works have explored the one-shot video inpainting problem, we compare our method to the one and only open-source one-shot video inpainting method, IIVI [28]. Furthermore, we also directly concatenate the state-of-the-art SVOS model STCN [8] with several state-of-the-art fully supervised video inpainting methods for comparison in the one-shot setting, including VInet [16], CAP [19], LGTSM [5], STTN [44], DSTT [23], FuseFormer [22], and E^{2}FGVI [21]. Specifically, we first use STCN to predict target mask of each frame in the videos except for the first frame, and then feed the predicted masks to traditional methods as target guidance. For a fair comparison, these predicted masks also undergo dilation operations before sending to the inpainting models following [44, 23, 22, 21]. Therefore, *all the above compared methods in our experiments are evaluated in the same one-shot setting as ours*.

Qualitative comparison. We first compare our method with four representative methods qualitatively. Here, we choose E^{2}FGVI [21] as the backbone of our method. Fig. 4 illustrates the qualitative results of different methods. We can see that all the traditional methods (Fig. 4(c)-(e)) suffer from severe residue artifacts. The reason behind this is that these methods are all trained with a hard masking mechanism on the ground truth masks, which gives rise to unsatisfactory results when the predicted target masks are inaccurate

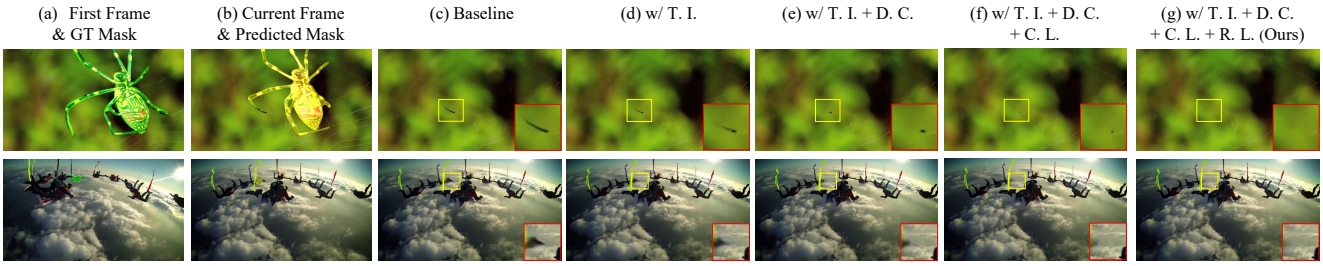

Figure 8: Qualitative comparison of ablations on the YouTube-VOS dataset. Best viewed with zooming in digital version.

during testing. Even worse, the residues of an inpainted frame may propagate to the subsequent frames, which further aggravates the residues issue. In addition, although IIVI [28] supports one-shot setting, it relies on a mask propagation mechanism that is sensitive to non-rigid deformation of the target, and thus it fails to capture the targets across frames with deformations (Fig. 4(h)). Even we modify IIVI to directly use the predicted masks by STCN [8] as a more accurate target guidance, the results are still not satisfactory (Fig. 4(f)). Unlike all the competitors, our method can create faithful and temporal consistent results due to the superior tolerance to inaccurate masks and delicate residue removal capability. Since our framework can be easily plugged into the existing video inpainting models. We also depict the qualitative results of different backbones integrated with our framework in Fig. 5. As can be observed, our framework renders consistent improvements in the performances of all the compared backbones. This demonstrates the huge potential of our method for practical applications.

Quantitative comparison. We also conduct a quantitative comparison with seven methods and report the results of our method based on different representative backbones, including DSTT [23], FuseFormer [22], and E^{2}FGVI [21]. In addition to comparing global metrics on the entire regions of images, we also report the local metrics by only computing the corresponding statistics of the target regions. The local metrics focus on measuring the inpainting quality of the target regions only, which facilitates a more intuitive comparison of the inpainting quality of different methods. As shown in Table 1, our method can effectively convert the traditional video inpainting methods to the one-shot setting, such that the predicted inaccurate masks can be properly tackled to achieve compelling inpainting results. This contributes to consistent improvements in all three quantitative metrics. The advantages of our method become more apparent on the local metrics, indicating the effectiveness of our method in producing plausible content in the target regions.

User study. We also conduct a user study to demonstrate the superiority of our method. Here, E^{2}FGVI [21] is chosen as our backbone. We select CAPnet [19], FuseFormer [22], and E^{2}FGVI [21] for comparison. In total, 20 real videos are selected for evaluation and 30 volunteers participate in the study. For each selected video, the volunteers are asked to choose the one with the best inpainting quality and temporal coherence among the results from different methods. The results are reported in Fig. 7, demonstrating the significant advantages of our method over the other methods in performing faithful and temporal-consistent inpainting.

Table 2: Ablations on the synthetic dataset. Inference time is measured on a 36-frame video with resolution 432 $\times$ 240. Online components are underlined.

Methods	Local				Inference Time (s)
	PSNR↑	SSIM↑	VFID↓	$\mathcal{J}\&\mathcal{F}$↑	
Baseline	28.11	0.865	0.764	94.84	9.18
w/ T. I.	28.77	0.879	0.755	94.84	9.22
w/ T. I. + D. C.	28.89	0.883	0.730	94.84	9.22
w/ T. I. + D. C. + C. L.	28.96	0.886	0.722	95.19	37.55
w/ T. I. + D. C. + C. L. + R. L. (Ours)	**29.01**	**0.888**	**0.719**	**95.32**	40.67

4.3. Ablation Study

In this section, we present an in-depth analysis of the effectiveness of our proposals, including the target inactivation module, dual-curriculum learning scheme, and residue removal module. We also report the time statistics of each model variant. E^{2}FGVI [21] is used as our backbone.

Effectiveness of Target Inactivation. To verify the efficacy of the target inactivation, we compare the inpainting results of the backbone (Baseline) and the backbone integrated with our target inactivation mechanism (w/ T. I.). As shown in Fig. 8(c), the results of the Baseline contain noticeable artifacts since it directly applies a hard predicted target mask to mask out the corresponding region of the input image, which inevitably ruins the meaningful image content and degrades the inpainting quality (Table 2). By contrast, our target inactivation mechanism allows better exploitation of the rich information in the target regions, by selectively inactivating the target responses in the latent space. This effectively benefits the overall learning and improves the inpainting results (Fig. 8(d) and Table 2).

Effectiveness of Dual-Curriculum Learning. We also explore the effectiveness of the dual-curriculum learning scheme (D. C.) by converting the target inactivation mechanism to a curricular one. As observed in Fig. 8(e), some

(a) long-form video

(b) high-resolution video

Figure 9: Results on long-form (973 frames) and high-resolution (1920 × 1080) videos. The first column denotes the first frames & GT masks and the rest columns denote the inpainted results. The red boxes represent the current frames & predicted masks.

undesired pixels outside the incompletely predicted target mask (*e.g.*, the legs of the spider in the first row) can be largely erased when the D. C. is adopted. This is because the D. C. effectively prevents the inpainting network from overly relying on the ground truth masks. This grants the model the capability to yield reasonable results even with erroneously predicted target masks, which is vital for reliable one-shot inpainting. Table 2 also verifies quantitative improvements contributed by the D. C., which are consistent with our observations in the qualitative results.

Effectiveness of Online Residue Removal. To investigate the significance of the online residue removal scheme, we analyze the gains brought by the contextual loss (C. L.) and the residue loss (R. L.) for fine-tuning the SVOS network during the inference phase. As can be seen in Fig. 8(f), the residues can be alleviated to some extent by adding the C. L., compared to the variant with no residue removal (Fig. 8(e)). After further incorporating the R. L., our full model finally produces visually appealing results with artifacts effectively removed (Fig. 8(g)). Table 2 also verifies both C. L. and R. L. boost the inpainting performance in terms of both inpainting quality and temporal coherence. It is also worth noting that the $\mathcal{J}\&\mathcal{F}$ metric gains prominent improvements (from 94.84 to 95.32) by adopting the two losses. This indicates that both losses can indeed improve prediction of target masks by the SVOS network, contributing to the final inpainting performance.

Time Statistics. Here we analyze the inference time of different variants of our method. As reported in Table 2, compared to the baseline, incorporating target inactivation or dual-curriculum learning incurs only negligible extra time costs. Due to the testing-time optimization, there is a noticeable increase in inference time when involving the online residue removal. However, our method can still bring significant performance improvements and achieve state-of-the-art even if the online reside removal is omitted.

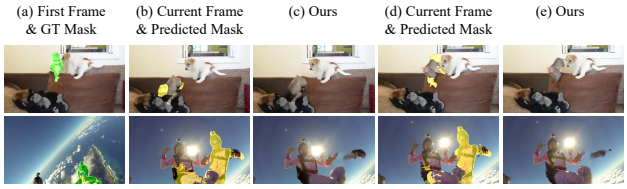

Figure 10: Failure cases of our framework.

4.4. Applicability Analysis of CIRI

To further verify the applicability of our CIRI to different scenarios, we also analyze its performance on long-form and high-resolution videos, respectively. As shown in Fig. 9(a), our framework can deliver consistently satisfactory results on long-form videos. Moreover, benefiting from the plug-and-play design, our method also performs fairly well on high-resolution videos by simply equipping with a high-resolution backbone network, *e.g.*, E^{2}FGVI-HQ [21](Fig. 9(b)). All the above results demonstrate the superior scalability and applicability of our method.

5. Limitations

Although our framework offers a simple yet effective solution to one-shot video inpainting problem. It may fail when the appearance of the target changes drastically across frames. As observed in Fig. 10, both the monkey and the pilot exhibit substantial appearance changes from the initial frame to the current frame. Consequently, the SVOS model produces inadequate segmentation masks, leading to incorrect target inactivation and ultimately resulting in inferior inpainting results.

6. Conclusion

In this paper, we propose a curricular inactivation framework for residue-aware one-shot video inpainting. Specifically, a curricular target inactivation module is first proposed to enhance the tolerance of the model to inaccurate target masks. Furthermore, we present an online residue removal scheme to effectively erase the residue artifacts during the testing phase. Extensive experiments demonstrate that our method can be readily plugged into the existing traditional inpainting methods to support one-shot setting with significant performance improvements.

Acknowledgements. The work is supported by Guangdong International Technology Cooperation Project (No. 2022A0505050009); China National Key R&D Program (Grant No. SQ2022YFE020322); Key-Area Research and Development Program of Guangzhou City (No. 2023B01J0022); National Natural Science Foundation of China (No. 61972162); Guangdong Natural Science Funds for Distinguished Young Scholars (No. 2023B1515020097); and Singapore Ministry of Education Academic Research Fund Tier 1 (MSS23C002).

References

[1] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, pages 119–135, 2018. 1

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009. 4

[3] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *ECCV*, pages 777–794, 2020. 3

[4] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017. 3

[5] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Learnable gated temporal shift module for deep video inpainting. In *BMVC*, 2019. 2, 7

[6] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *CVPR*, pages 9384–9393, 2020. 3

[7] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, pages 1189–1198, 2018. 3

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 34:11781–11794, 2021. 3, 6, 7, 8

[9] Mounira Ebdelli, Olivier Le Meur, and Christine Guillemot. Video inpainting with short-term windows: application to object removal and error concealment. *IEEE TIP*, 24:3034–3047, 2015. 2

[10] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, pages 713–729, 2020. 2

[11] Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *ECCV*, pages 682–695, 2012. 2

[12] Yu Guo, Yuan Gao, Wen Liu, Yuxu Lu, Jingxiang Qu, Shengfeng He, and Wenqi Ren. Scanet: Self-paced semi-curricular attention network for non-homogeneous image dehazing. In *CVPR Workshop*, pages 1884–1893, 2023. 4

[13] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, pages 54–70, 2018. 3

[14] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM TOG*, 38:1–13, 2019. 1

[15] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In *CVPR Workshop*, 2017. 3

[16] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *CVPR*, pages 5792–5801, 2019. 2, 7

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 7

[18] Thuc Trinh Le, Andrés Almansa, Yann Gousseau, and Simon Masnou. Object removal from complex videos using a few annotations. 5:267–291, 2019. 1, 2, 3

[19] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *ICCV*, pages 4413–4421, 2019. 2, 7, 8

[20] Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, and C-C Jay Kuo. Depth-aware stereo video retargeting. In *CVPR*, pages 6517–6525, 2018. 1

[21] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *CVPR*, pages 17562–17571, 2022. 2, 5, 6, 7, 8, 9

[22] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, pages 14040–14049, 2021. 2, 5, 6, 7, 8

[23] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, and Li Hongsheng. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021. 2, 7, 8

[24] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, pages 768–783, 2018. 5

[25] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, pages 7376–7385, 2018. 3

[26] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 3

[27] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *ICCV*, pages 4403–4412, 2019. 2

[28] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *ICCV*, pages 14579–14588, 2021. 2, 3, 7, 8

[29] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 2, 7

[30] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *IJCV*, 2022. 2, 7

[31] Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *CVPR*, pages 3511–3520, 2022. 2

[32] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. In *ECCV*, pages 212–228, 2020. 4

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014. 5

[34] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast

end-to-end embedding learning for video object segmentation. In *CVPR*, pages 9481–9490, 2019. 3

[35] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *BMVC*, 2017. 3

[36] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *CVPR*, pages 17694–17703, 2022. 1

[37] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, pages 1328–1338, 2019. 5, 6

[38] Zhiliang Wu, Hanyu Xuan, Changchang Sun, Kang Zhang, and Yan Yan. Semi-supervised video inpainting with cycle consistency constraints. *arXiv preprint arXiv:2208.06807*, 2022. 2

[39] Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, and Shengfeng He. Towards a smaller student: Capacity dynamic distillation for efficient image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015, 2023. 6

[40] Cheng Xu, Wei Qu, Xuemiao Xu, and Xueting Liu. Multi-scale flow-based occluding effect and content separation for cartoon animations. *IEEE TVCG*, 2022. 2

[41] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018. 2, 6, 7

[42] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *CVPR*, pages 3723–3732, 2019. 2

[43] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, pages 6499–6507, 2018. 3

[44] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, pages 528–543, 2020. 2, 6, 7

[45] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *ICCV*, pages 2720–2729, 2019. 2

[46] Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, and Yong Du. Curricular contrastive regularization for physics-aware single image dehazing. In *CVPR*, pages 5785–5794, 2023. 4

[47] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *CVPR*, pages 16448–16457, 2021. 2