Occlusion-Aware Video Object Inpainting

Lei Ke[1]   Yu-Wing Tai[2]   Chi-Keung Tang[1]

[1]The Hong Kong University of Science and Technology   [2]Kuaishou Technology

{lkeab,cktang}@cse.ust.hk, yuwing@gmail.com

Abstract

Conventional video inpainting is neither object-oriented nor occlusion-aware, making it liable to obvious artifacts when large occluded object regions are inpainted. This paper presents occlusion-aware video object inpainting, which recovers both the complete shape and appearance for occluded objects in videos given their visible mask segmentation.

To facilitate this new research, we construct the first large-scale video object inpainting benchmark YouTube-VOI *to provide realistic occlusion scenarios with both occluded and visible object masks available. Our technical contribution VOIN jointly performs video object shape completion and occluded texture generation. In particular, the shape completion module models long-range object coherence while the flow completion module recovers accurate flow with sharp motion boundary, for propagating temporally-consistent texture to the same moving object across frames. For more realistic results, VOIN is optimized using both T-PatchGAN and a new spatio-temporal attention-based multi-class discriminator.*

Finally, we compare VOIN and strong baselines on YouTube-VOI. Experimental results clearly demonstrate the efficacy of our method including inpainting complex and dynamic objects. VOIN degrades gracefully with inaccurate input visible mask.

1. Introduction

Conventional video inpainting infers missing pixel regions by distilling information from remaining unmasked video regions. However, as shown in Figure 1, these models [1, 2, 3, 4] often fail to recover moving objects with large occlusion by wrongly inpainting the occluded region with irrelevant background colors and produce obvious artifacts. This is due to their lack of object and occlusion awareness. On the contrary, our human visual system possesses powerful amodal perception ability to reason the complete structure of moving objects under occlusion, including the appearance of the invisible regions in high fidelity [5, 6].

To overcome the above limitations, we make the first significant attempt on occlusion-aware video object inpainting, which completes occluded video objects by recovering their shape and appearance in motion. While there exist object completion models, they are applicable only to single images [7] in highly limited scenarios such as car and indoor furniture [8, 9]. These single-image models do not utilize temporal coherence during mask and content generation, thus leading to temporal artifacts and unsmooth transition when directly applied to videos.

Training models for occlusion reasoning requires a large number and variety of occluded video objects with amodal mask annotations. One difficulty arises from the existing amodal datasets, which come mostly in single images and are small [11, 12], or cover very limited object classes [13]. Inspired by [7], our approach is trained in a self-supervised manner with only modal annotations. To create realistic data in large quantity, we contribute the first large-scale video object inpainting benchmark with diverse occlusion patterns and object classes for both training and evaluation. Specifically, we generate occlusion masks for video objects by high-fidelity simulation of overlapping objects in motion, thus taking into consideration object-like occlusion patterns, motion and deformation under various degrees of occlusion. Our new **YouTube-VOI** dataset based on YouTube-VOS [14] contains 5,305 videos, a 65-category label set including common objects such as people, animals and vehicles, with over 2 million occluded and visible masks for moving video objects.

To infer invisible occluded object regions, we propose the **VOIN** (**V**ideo **O**bject **I**npainting **N**etwork), a unified multi-task framework for joint video object mask completion and object appearance recovery. Our object shape completion module learns to infer complete object shapes from only visible mask regions and object semantics, while our appearance recovery module inpaints occluded object regions with plausible content. To obtain pixel-level temporal coherency, we design a novel occlusion-aware flow completion module to capture moving video object and propagate consistent video content across even temporally distant frames, by enforcing flow consistency during the in-

---

| Corrupted Input | LGTSM | FGVC | STTN | Ours |

Figure 1. Video object inpainting results comparison with state-of-the-art LGTSM [10], FGVC [2] and STTN [3]. Our VOIN takes corrupted video with free-form occlusion masks as input, and faithfully recovers the occluded object region while preserving spatial detail and temporal coherence.

visible object region generation. Furthermore, we augment the temporal patch-based GAN training process using a new multi-class discriminator with specially designed spatio-temporal attention module (STAM), which effectively accelerates model convergence and further improves inpainting quality.

Finally, we evaluate VOIN and strong adapted baselines on *YouTube-VOI* benchmark, where quantitative and qualitative results clearly demonstrate VOIN's advantages. This paper complements conventional video inpainting and facilitates future development of new algorithms on repairing occluded video objects.

2. Related Work

Video Inpainting. Previous works [15, 16, 17, 18, 19] on video inpainting fill arbitrary missing regions with visually pleasing content by learning spatial and temporal coherence, with deep learning based approaches [20, 4, 21, 22, 23, 24, 3, 1, 10] becoming mainstream in recent years. The first deep generative model applied in video inpainting [20] combines 3D and 2D convolutions to produce temporally consistent inpainting content. To achieve better temporal consistency, in [4, 21, 2] optical flows are used to guide propagation of information across frames. In [22] a temporal memory module is used with recurrent feedback. For modeling long-range dependencies, in [23] frame-wise attention is applied on frames aligned by global affine transformation, and in [24] the authors adopt pixel-wise attention to progressively fill the hole from its boundary. Temporal PatchGAN [1] based on SN-PatchGAN [25] and temporal shift modules [10] are proposed to further enhance inpainting quality. Most recently, a spatio-temporal transformer is proposed in [3] for video completion by using a multi-scale patch-based attention module.

Although promising results have been achieved by the above methods, they still cannot satisfactorily recover the appearance of occluded video objects due to their lack of object and occlusion awareness. On the other hand, our VOIN jointly optimizes the *amodal* shape and flow completion for occluded objects. Different from previous flow completion [4, 21, 2] for random missing areas, our flow completion can faithfully recover the flow within occluded

object regions. Using our large-scale synthetic data with accurate flow available for training, we can enforce and maintain strong flow consistency in spatial and temporal texture generation not previously possible in other methods. Together with the predicted amodal mask, we can thus produce more accurate object flow with sharper motion boundary than [4, 2]. Moreover, semantic information is incorporated into our proposed spatio-temporal multi-class discriminator, which makes the GAN training process faster and more stable, and further enhances the inpainting quality for unseen regions.

Amodal Object Completion. In amodal object completion, visible masks of objects are given and the task is to complete the modal into amodal masks, which is different from amodal instance segmentation [26, 11, 27, 13]. Previous amodal mask completion approaches make assumptions about the occluded regions, such as Euler Spiral [28], cubic Béziers [29] and simple curves (straight lines and parabolas) [30]. These unsupervised methods cannot deal with objects with complex shapes.

Occlusion handling has also been extensively studied [31, 32, 33, 34, 35, 36, 37, 38, 39], especially in object detection [40, 41, 42, 43], segmentation [44] and tracking [45, 46, 47, 48], but most do not consider recovering the appearance of the occluded objects. Among the prior amodal object completion works with appearance recovery, Ehsani *et al* [8] generate the occluded parts of objects using Unet [49] by leveraging about 5,000 synthetic images restricted to indoor scenes such as kitchen and living room. Yan *et al* [9] recover the appearance of occluded cars by synthesizing occluded vehicle dataset. Zhan *et al* [7] propose a self-supervised scene de-occlusion method PCNet which can complete the mask and content for invisible parts of more common objects without amodal annotations as supervisions. However, all of these methods are single image-based without considering temporal coherence and object motions. Extending them directly to complex video sequence may easily lead to unwanted temporal artifacts.

3. Video Object Inpainting Network

Given $\mathbf{X}_1^T = \{X_1, X_2, ..., X_T\}$ as an input video sequence with frame length T, frame resolution $H \times W$,
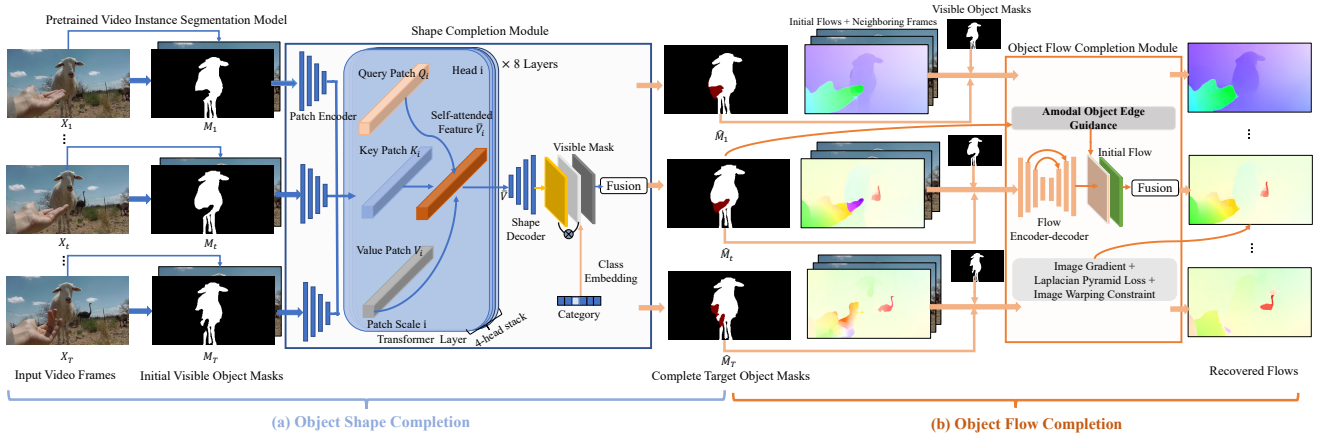
Figure 2. **(a)** Object shape completion, which associates transformed temporal patches and object semantics; **(b)** Object flow completion, which recovers complete object flow subject to the amodal object contours. Red indicates occluded region.

and $\mathbf{M}_1^T = \{M_1, M_2, ..., M_T\}$ denotes the corresponding frame-wise binary masks for the visible regions of the target occluded object, we formulate the video object inpainting problem as self-supervised learning to infer complete object masks $\hat{\mathbf{M}}_1^T = \{\hat{M}_1, \hat{M}_2, ..., \hat{M}_T\}$ and produce completed video frames $\mathbf{Y}_1^T = \{Y_1, Y_2, ..., Y_T\}$ with realistic amodal object content.

Figures 2 and 3 together depict the whole pipeline of our proposed video object inpainting approach VOIN, which consists of the following the three stages: a) *object shape completion*: we compute the amodal object shapes based on its visible object content (Section 3.1); b) *object flow completion*: complete object flow is estimated with sharp motion boundary under the guidance of amodal object contour (Section 3.2); c) *flow-guided video object inpainting*: with the completed object and flow within its contour, motion trajectories are utilized to warp pertinent pixels to inpaint the corrupted frames. To generate highly plausible video content, we improve temporal shift module by making it occlusion-aware and use multi-class discriminator with spatio-temporal attention, instead of only using single-image completion techniques as in [2, 4] (Section 3.3).

3.1. Occlusion-Aware Shape Completion

Given an input video sequence, it is easy to obtain the modal masks for the target occluded object using the existing video object/instance segmentation [50, 51, 52]. However, learning the full completion video masks of occluded instances is very difficult due to diverse object shapes and occlusion patterns.

To address this problem, we propose a novel object shape completion module (see Figure 2(a)), which recovers amodal segmentation masks for the occluded video object in a self-supervised training scheme. Our shape module with 8 transformer layers is inspired from the recent spatio-temporal transformers in video understanding [53, 54, 55, 3, 56] for capturing long-range spatio-temporal coherence. Each transformer layer has a multi-head structure to deal with the multi-scaled embedded image patches transformed from the whole input video sequence, followed by the scaled dot-product attention mechanism [57], which

models temporal shape associations of the same occluded object among both the neighboring and distant encoded spatial feature patches in parallel.

Specifically, suppose there are k heads in the transformer layer, then we compute self-attended feature $\bar{V}$:

$$\bar{V} = \text{Multihead}(Q, K, V) = f_c([\bar{V}_i]_{i=1}^k), \qquad (1)$$

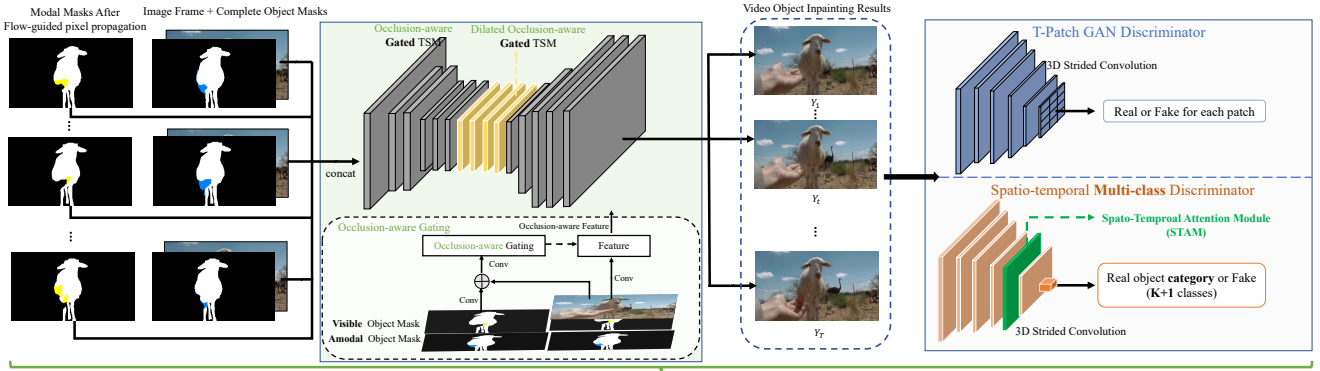$$\bar{V}_i = \text{SelfAtt}(Q_i, K_i, V_i) = \frac{softmax(Q_i K_i^T)}{\sqrt{d_k}} V_i, \qquad (2)$$

where $\bar{V}_i$ is the self-attended feature on the i-th head, Q_i, K_i, V_i are respectively the query, key and value embedding matrices for these spatial feature patches with total size $T \times H/r_1 \times W/r_2$, frame resolution is $H \times W$, r_1 and r_2 are patch size, d_k is the dimension for query patch features, and f_c are convolutional layers merging the outputs from the k heads. $\bar{V}$ is then passed through the frame-level shape decoder for up-sampling, which is combined with the class embedding feature multiplied by the visible object masks for incorporating semantics and spatial shape prior. Finally, the merged features are refined by the fusion convolution layers to produce the amodal object shape masks.

3.2. Occlusion-Aware Flow Completion

Our flow completion algorithm first computes the initial optical flows and then focuses on recovering the flow fields within the completed occluded object region subject to the amodal object contour.

In Figure 2(b), the flow generator adopts Unet [49] encoder-decoder structure with skip connections from encoders to the corresponding layers in the decoder, which takes neighboring image frames, initial flow, visible and amodal object masks as inputs x. Instead of computing the recovered flow directly, we formulate flow completion as a residual learning problem [58], where $\phi(x) := O - \bar{O}$, O is the desired flow output, $\bar{O}$ is the initial corrupted flow, and $\phi(x)$ represents the flow residue learned by the encoder-decoder generator. This formulation effectively reduces the training difficulty for dense pixel regression.

To recover accurate object flow with sharp motion boundary, especially for the occluded region, we incorpo-

(c) Flow-guided Video Object Inpainting

Figure 3. **(c)** Flow-guided video object inpainting with the occlusion-aware gating scheme. Figure 4 further illustrates the operation of our occlusion-aware temporal shift module (TSM). We adopt T-patch Discriminator [1, 25] and propose a multi-class discriminator with spatio-temporal attention module (STAM) to regularize the GAN training. STAM is detailed in Figure 5. Yellow marks the region filled by known pixels tracked by optical flow; blue regions are unseen/occluded regions to be inpainted in the final stage.

rate amodal object contour to guide the flow prediction process by enforcing flow smoothness within the complete object region, as flow fields are typically piecewise smooth, where gradients are small except along the distinct object motion boundaries. To effectively regularize the flow completion network, instead of simply adopting L1 regression loss between predictions and ground-truth flows as in [4], we additionally utilize the image gradient loss, Laplacian Pyramid loss [59] and image warping loss of hallucinated content for joint optimization, which further promote the precision of flow prediction and are detailed in section 3.4.

3.3. Flow-Guided Video Object Inpainting

The resulting completed object flow above is employed to build dense pixels correspondences across frames, which is essential since previously occluded regions in the current frame may be disoccluded and become visible in a distant frame, especially for objects in slow motion, which is very difficult for a generative model to handle such long-range temporal dependencies.

We follow [4, 2] using forward-backward cycle consistency threshold (5 pixels) to filter out unreliable flow estimations and warp pixels bidirectionally to fill the missing regions based on the valid flow. The main difference is that we only warp pixels within the foreground object region, which guarantees the occluded areas are not filled by any background colors while reducing the overall computational burden. Figure 3 highlights the regions in yellow within a completed mask tracked by optical flow.

To fill in remaining pixels after the above propagation (i.e., the blue regions in Figure 3), which can be in large numbers for previously heavily occluded objects, we propose to train an occlusion-aware gated generator to inpaint the occluded regions of videos objects, where the gating feature is learned under the guidance of both amodal object masks and occlusion masks, and two spatio-temporal discriminators with multi-class adversarial losses. As usual, the discriminators will be discarded during testing.

3.3.1 Occlusion-Aware TSM

We adopt the residual Temporal Shift Module (TSM) [60, 10] as our building blocks here, which shift partial channels along the temporal dimension to perform joint spatio-temporal feature learning, and achieve the performance of 3D convolution at 2D CNN's complexity.

However, the original TSM treats all feature points equally, making no difference between visible and occluded regions belonging to the same object, and failing to distinguish invalid feature points from corrupted hole areas. Thus, to make TSM occlusion-aware and learn a dynamic feature selection mechanism for different spatial locations, we guide the gated feature learning process [25] with both amodal object masks and occlusion masks, thus making our improved or occlusion-aware model capable of reasoning the occluded regions from other visible parts along the spatio-temporal dimension as illustrated in Figure 4.

Specifically, the generator in Figure 3 has the encoder-decoder structure with Occlusion-aware TSM replacing all vanilla convolutions layers, which have a larger temporal receptive field n than original setting [60] and can be formulated as

$$Gate_{\text{occ}}^{x,y}(t) = \sum_{x,y} W_g \cdot I_t^{x,y} + \bar{f}_t^{x,y}(\hat{M}_t^{occ}, \hat{M}_t), \quad (3)$$

$$\mathcal{S}_t^{x,y} = \sum_{x,y} W_f \cdot TSM(I_{t-n}^{x,y}, \cdots, I_t^{x,y}, \cdots, I_{t+n}^{x,y}), \quad (4)$$

$$Out_t^{x,y} = \sigma(Gate_{\text{occ}}^{x,y}(t)) \odot \phi(\mathcal{S}_t^{x,y}(t)), \quad (5)$$

where $Gate_{occ}^{x,y}$ serves as a soft attention map (for identifying occluded/visible/background areas) on the feature volume $\mathcal{S}_t^{x,y}$ output by the TSM, $\bar{f}$ are convolutional layers fusing the occlusion mask $\hat{M}_t^{occ}$ and complete object mask $\hat{M}_t$, W_g and W_f are respectively the kernel weights for gating convolution and shift module, and $I_t^{x,y}$ and $Out_t^{x,y}$ respectively denote the input and final output activation at (t, x, y), σ is sigmoid function, and ϕ is the ReLU function.

3.3.2 Multi-Class Discriminator with STAM

To make object inpainting results more realistic, we adopt two discriminators to simultaneously regularize the GAN
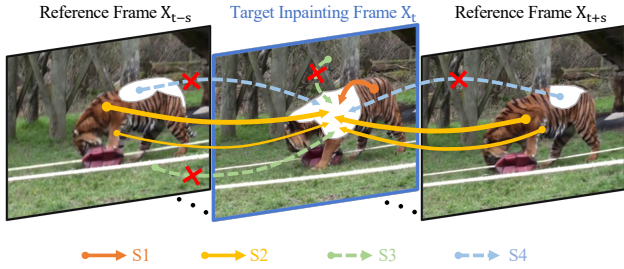
Figure 4. Illustration of our occlusion-aware gating scheme. To inpaint the occluded region (in white) of the tiger in target frame X_t, VOIN only learns from valid visible object parts along spatio-temporal dimension indicated by S1 and S2, while excluding irrelevant background pixels or occluded pixels by S3 and S4.

training process. The first discriminator considers video perceptual quality and temporal consistency, while the second considers object semantics based on both the global and local features since the occlusion holes may appear anywhere in the video with irregular shape.

We adopt T-PatchGAN as the first discriminator [1, 3]. For the second discriminator, we propose a new spatio-temporal attention-based multi-class discriminator, which classifies the category of the inpainting object into one of the K real classes and an additional fake class, by picking the most relevant frames from an input video while focusing on their discriminative spatial regions. Figure 3 shows that the multi-class discriminator is composed of six 3D convolution layers (kernel size $3 \times 5 \times 5$) with a spatio-temporal attention module (STAM) embedded above the 4th layer. This STAM design is inspired by [61, 62]. Figure 5 shows the parallel branches for spatial and temporal attention.
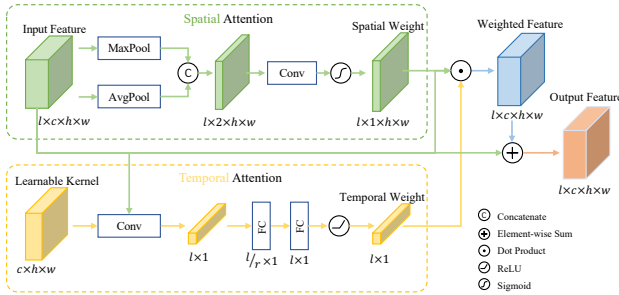

Figure 5. The design of spatio-temporal attention module (STAM), where two parallel branches are respectively used for computing spatial and temporal attention weights. The weighted feature forms a residual connection with the original input for final output.

Although the deep layers of T-PatchGAN [1] can cover whole videos with its large receptive field, we find the multi-class discriminator augmented with object semantics and STAM effectively accelerates model convergence and further improves inpainting quality (see Table 4).

3.4. Optimization Objectives

We train the multi-task VOIN framework in joint optimization with comprehensive objectives designed to produce sharp and spatio-temporal conherent video content,

which are respectively shape completion loss $\mathcal{L}_{\text{shape}}$, flow completion loss $\mathcal{L}_{\text{flow}}$, and appearance recovery loss $\mathcal{L}_{\text{app}}$.

For video object shape completion, in addition to the traditional binary cross entropy (BCE) loss, we also adopt dice loss [63] to resolve the imbalance between the number of foreground and background pixels, since the occluded regions sometimes only occupy a small area in the whole image. Thus L_{shape} is formulated as

$$\mathcal{L}_{\text{shape}} = \mathcal{L}_{\text{BCE}}(M', \hat{M}) + \lambda_1 \mathcal{L}_{\text{Dice}}(M'_{\text{occ}}, \hat{M}_{\text{occ}}), \quad (6)$$

where M' and M'_{occ} respectively denote the predicted complete masks and the masks for occluded region, $\hat{M}$ and $\hat{M}_{\text{occ}}$ are respectively the corresponding ground truth masks, and λ_1 is the balance weight.

For dense object flow completion, we enforce the recovered flow with both the pixel-level accuracy and smooth flow field, where $\mathcal{L}_{\text{flow}}$ is designed as

$$\mathcal{L}_{\text{flow}} = ||(1+\hat{M}) \odot (\hat{O}-O)||_1 + \lambda_2 \mathcal{L}_{\text{Lap}}(O, \hat{O}) + \mathcal{L}_g + \mathcal{L}_w, \quad (7)$$

$$\mathcal{L}_g = \hat{M} \odot (\lambda_3 ||(G_\Delta(O) - G_\Delta(\hat{O}))||_1 + \lambda_4 ||G_\Delta(O)||_1), \quad (8)$$

where $\hat{O}$ and O are respectively the predicted and ground-truth flow, G_Δ computes flow gradients in both horizontal and vertical directions, where we use $\mathcal{L}_g$ to minimize the gradients for non-edge pixels inside the complete object to ensure smooth continuation while keeping object motion boundary with sharp transitions. $\mathcal{L}_{\text{Lap}}$ is used for preserving details at different spatial scales [64], and $\mathcal{L}_w$ supervises image warping consistency using the predicted flow.

To recover plausible object appearance, we optimize both the binary T-PatchGAN discriminator D_p (to discriminate real or fake content) [1] and our multi-class global discriminator D_{cls} (to classify category, with K real classes and one fake class) using the spatio-temporal adversarial loss and semantic loss. The optimization function $\mathcal{L}_{\text{Dis}}$ for the two discriminators is defined as

$$\mathcal{L}_{\text{Dis}} = \mathbb{E}_{x \sim p_{\text{data}}(x)}[1 - D_p(x)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(D_{\text{cls}}(y|x))] \\ + \mathbb{E}_{z \sim p_z(z)}[1 + D_p(G(z))] + \mathbb{E}_{z \sim P_z(z)}[\log(D_{\text{cls}}((K+1)|G(z)))], \quad (9)$$

where $y \in \{1, ..., K\}$, $ReLU$ is omitted for simplicity and the loss $\mathcal{L}_{\text{Gen}}$ for inpainting generator is

$$\mathcal{L}_{\text{Gen}} = -\mathbb{E}_{z \sim p_z(z)}[D_p(G(z))] - \mathbb{E}_{z \sim p_z(z)}[D_{cls}((K+1)|G(z))], \quad (10)$$

The per-pixel content reconstruction loss $\mathcal{L}_{\text{content}}$ and appearance recovery loss $\mathcal{L}_{\text{app}}$ are respectively defined as

$$\mathcal{L}_{\text{content}} = ||\hat{M} \odot (Y' - Y)||_1 + \lambda_5 ||(1-\hat{M}) \odot (Y' - Y)||_1, \quad (11)$$

$$\mathcal{L}_{\text{app}} = \mathcal{L}_{\text{Dis}} + \mathcal{L}_{\text{Gen}} + \lambda_6 \mathcal{L}_{\text{content}}, \quad (12)$$

where Y' and Y are respectively the ground truth and predicted completed frame. Thus, the overall optimization objectives are summarized as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{shape}} + \lambda_{\text{flow}} \mathcal{L}_{\text{flow}} + \lambda_{\text{app}} \mathcal{L}_{\text{app}}. \quad (13)$$

4. Experiments

4.1. Dataset and Evaluation Metrics

Occlusion Inpainting Setting Since this paper focuses on inpainting occluded regions of video objects, we propose a new inpainting setting which is different from previous ones for undesired object removal or arbitrary mask region inpainting. The fill-up regions are restricted to occluded regions of the target object, which can be given by user or our object shape completion module using the visible object content. This setting is in line with real-world applications such as video scene de-occlusion (Figure 11).

YouTube-VOI Benchmark. To support training and evaluation of our new video object inpainting task, we use the YouTube-VOS [14] dataset as our video source to construct our large-scale YouTube-VOI benchmark, which contains 5,305 videos (4,774 for training and 531 for evaluation) with resolution higher than 640 × 480, a 65-category label set including common objects such as people, animals and vehicles, and over 2 million occluded and visible masks.
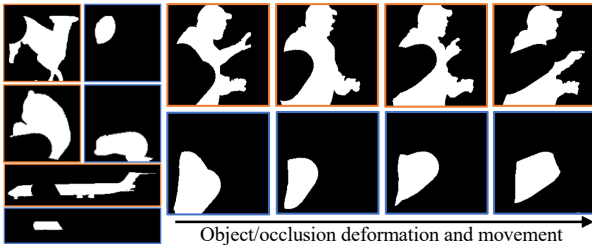

Figure 6. Sample visible masks (orange boxes) and occluded masks (blue boxes) for moving video objects generated by our algorithm with different object categories and occlusion patterns.

We generate both the occlusion masks and visible masks for video objects by high-fidelity simulation of overlapping objects in motions (Figure 6), where we take into consideration various object-like occlusion patterns, occluder movements and shape deformations under various degrees of occlusion from 10% to 70%.

Our YouTube-VOI is a very challenging dataset for video object inpainting, which is representative of complex real-world scenarios in high diversity, including different realistic occlusions caused by vehicles, animals and human activities. Although amodal object masks are not annotated in Youtube-VOS, we show that the proposed VOIN model can still perform amodal object shape completion for video with only modal annotations, by incorporating the self-supervised training scheme in [7] to learn object shape associations between frames, utilizing the annotated object semantics, and training on huge number of occlusions masks in various degrees and patterns of occlusion.

Evaluation Metrics. For evaluating video object inpainting quality, we use the widely adopted PSNR, SSIM and LPIPS metrics following [2, 4], where LPIPS [65] is obtained using Alexnet [66] as backbone with linear calibration on top of intermediate features as its default setting.

Implementation Details. We build our occlusion-aware inpainting generator adapted from the encoder-decoder

structure in [10]. For data preprocessing, we resize video frames to 384 × 216 and randomly crop them to 320 × 180 with random horizontal flip. We generate video object occlusion masks based on the free-from masks [25, 1]. For more network and implementation details, please refer to the supplementary materials.

4.2. Comparison with State-of-the-arts

Using the Youtube-VOI benchmark, we compare VOIN with the most recent and relevant state-of-the-art video inpainting approaches and adapt their original input by additionally concatenating visible object masks: 1) DFVI [4], which fills corrupted regions using pixel propagation based on the predicted complete flow; 2) LGTSM [10], where learnable shift module is designed for inpainting generator and T-PatchGAN discriminator [1] is utilized; 3) FGVC [2], which uses a flow completion module guided by Canny edge extraction [67] and connection [68]; 4) STTN [3], which completes missing regions using multi-scale patch-based attention module. Note that both DFVI and FGVC conduct flow completion for inpainting, and fill the remaining unseen video regions using only image inpainting method [69].

Flow Completion Comparison. We compare our object flow completion module with [4, 2] both qualitatively and quantitatively. Table 1 shows the endpoint error (EPE) between the pseudo ground truth flow computed from the original, un-occluded videos using RAFT [70] and the predicted completed flow on the missing regions. The quantitative result reveals that the proposed object flow completion module achieves significantly lower EPE errors than the previous flow completion networks in [4, 2]. Figure 7 compares flow completion results, where our proposed occlusion-aware flow completion module produces sharp motion boundary with a smooth flow field within the object contour, which shows the effectiveness of amodal object shape guidance and hybrid loss optimization. Although FGVC [2] trained a separate flow edge connection network using EdgeConnect [68], their model still cannot complete large occlusion holes but generates blurred and ambiguous flow completion result.

Table 1. Quantitative comparison on flow completion (EPE) and inpainting quality (PSNR, SSIM and LPIPS) on Youtube-VOI benchmark with the state-of-the-art methods.

Model	Use flow?	EPE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DFVI [4]	✓	4.79	44.91	0.952	0.099
LGTSM [10]		-	45.19	0.979	0.024
FGVC [2]	✓	3.69	43.90	0.924	0.065
STTN [3]		-	45.97	0.986	0.020
Ours		-	46.33	0.989	0.013
Ours	✓	**3.11**	**48.99**	**0.994**	**0.008**

Quantitative Results Comparison. Table 1 also reports quantitative comparison on inpainting quality under complex occlusion scenarios on the Youtube-VOI test set. Compared to existing models, our VOIN substantially improves
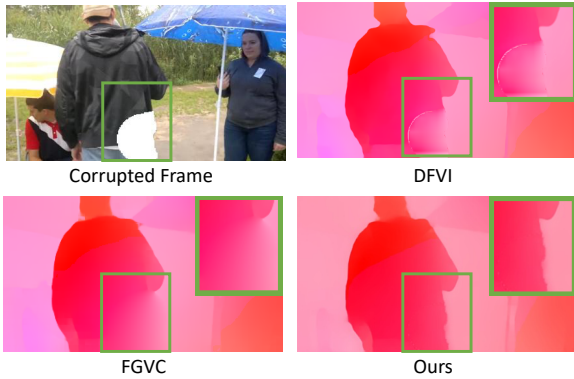
Figure 7. Flow completion results comparison with DFVI [4] and FGVC [2]. Our method completes object flow with sharper motion boundary and more natural piece-wise smooth transition with no visible seams within the hole compared to DFVI.

video reconstruction quality with both per-pixel and overall perceptual measurements, where our model outperforms the most recent STTN [3] and FGVC [2] by a large margin, especially in terms of PSNR and LPIPS. Our improved results show the effectiveness of our proposed occlusion-aware gating scheme and the multi-class discriminator with STAM. On the other hand, the results produced by DFVI and FGVC are not on par with ours, especially for inpainting foreground object with large occlusion due to their incorrect flow completion and the lack of temporal consistency for generating video content (limited in using single image-based inpainting model DeepFill [69]).

Qualitative Results Comparison. Figure 8 shows sample video completion results for inpainting occluded video objects, where our occlusion-aware VOIN produces temporally coherent and visually plausible content than previous methods [10, 2, 3]. Figure 11 shows its application in video scene de-occlusion. Refer to the supplementary video results for extensive qualitative comparison.

4.3. Ablation Study

Video Object Shape Completion. To evaluate the effectiveness of different components in occlusion-aware object shape completion module, Table 2 reports the corresponding ablation study by reducing transformer layers to 4 and removing semantic embedding and dice loss [63]. The spatial shape prior of object categories can greatly improve the mIoU by around 5%, and the dice loss further promotes the completion performance on small occluded regions. We also compare our video shape completion module with the image-based UNet trained on amodal annotation [7], where the large performance gap reveals the importance of learning object shape associations between frames.

Influence of Initial Mask Segmentation. Figure 9 shows the influence of initial segmentation quality on the final object inpainting results in the presence of mask errors. Please refer to Section 4 of the supplemental file for full mask degradation experimental results and analysis.

Video Object Flow Completion. Table 3 depicts the ablation experiment result on the proposed object flow completion module. We find that the end-to-end point flow er-

Table 2. Ablation study on our *object shape completion* module. B_{S-n}: video shape completion module with n transformer layers and BCE loss. D: using dice loss. S: adding semantic guidance.

	Image-based	B_{S-4}	B_{S-8}	B_{S-8} + D	B_{S-8} + D + S
mIoU (%)	74.81	78.02	80.28	81.53	**86.68**

ror is reduced progressively by successively incorporating amodal shape guidance and hybrid loss constraints. Also, we show the flow completion comparison in Figure 10 to visualize the effectiveness of each component. The amodal shape guidance prevents corruption due to irrelevant background flows, and the image gradient loss effectively retains the natural smooth flow transition within the completed object, although we observe that using gradient loss takes the flow completion network longer time to converge.

Table 3. Ablation study on our *object flow completion* module. B_F: flow completion baseline using Unet structure for regressing flow directly. A: amodal object edge guidance. L: Gradient loss and image warping loss.

	B_F	B_F + A	B_F + A + L
Flow completion (EPE)	4.89	3.95	**3.11**

Table 4. Ablation study on different VOIN components. B_I: VOIN baseline using the generator network [10]. OG: our occlusion-aware gating scheme. TP: T-PatchGAN discriminator [1]. MD: multi-class discriminator. STAM: our spatio-temporal attention module. F: our object flow completion guidance.

Model	Use flow?	PSNR ↑	SSIM ↑	LPIPS ↓
B_I		44.75	0.979	0.025
B_I + OG		45.25	0.982	0.019
B_I + OG + TP		45.54	0.984	0.017
B_I + OG + TP + MD		45.91	0.986	0.014
B_I + OG + TP + MD + STAM		46.33	0.989	0.013
B_I + OG + STAM	✓	48.16	0.991	0.012
B_I + OG + TP + MD + STAM	✓	**48.99**	**0.994**	**0.008**

4.3.1 VOIN Model

To investigate how each component in VOIN contributes to the final video object inpainting performance, especially for the proposed occlusion-aware gating scheme and multi-class discriminator with spatio-temporal attention module (STAM), Table 4 reports the ablation study results, and we analyze the effectiveness of our design choices as follows.

Effect of Occlusion-Aware Gating. Table 4 shows the importance of identifying occluded/visible/background areas on the input feature volume (illustrated in Figure 4), where our occlusion-aware gating scheme learns to infer the occluded object regions by attending to the visible spatio-temporal object regions without adversely affected by the background. This strategy significantly improves the perceptual similarity metric LPIPS by 24% and improves the PSNR from 44.75 to 45.25.

Effect of MD with STAM. Multi-class discriminator (MD) incorporates semantics into GAN training, which enables more fine-grained object classification and thus produces more realistic video content, improving PSNR from 45.54 to 45.91. Furthermore, STAM enhances the overall performance by enabling the multi-class discriminator to focus on more discriminative feature region through learning spatio-temporal attention weights across video frames,

| Corrupted Input | LGTSM | FGVC | STTN | Ours |

Figure 8. Qualitative comparison with state-of-the-art video inpainting methods LGTSM [10], FGVC [2], STTN [3] on Youtube-VOI. In particular, FGVC also adopts completed flow to guide the video inpainting process, but their results suffer from unnatural pixel transition due to the incorrect flow estimation. Zoom in for better view. Refer to the supplementary file for more qualitative comparison.

| Input frame | Detected mask of the occludee | Inpainting results using detected masks | Dilation for 14 pixels (deviation rate 25%) | Results after 14-pixel dilation (deviation rate 25%) |

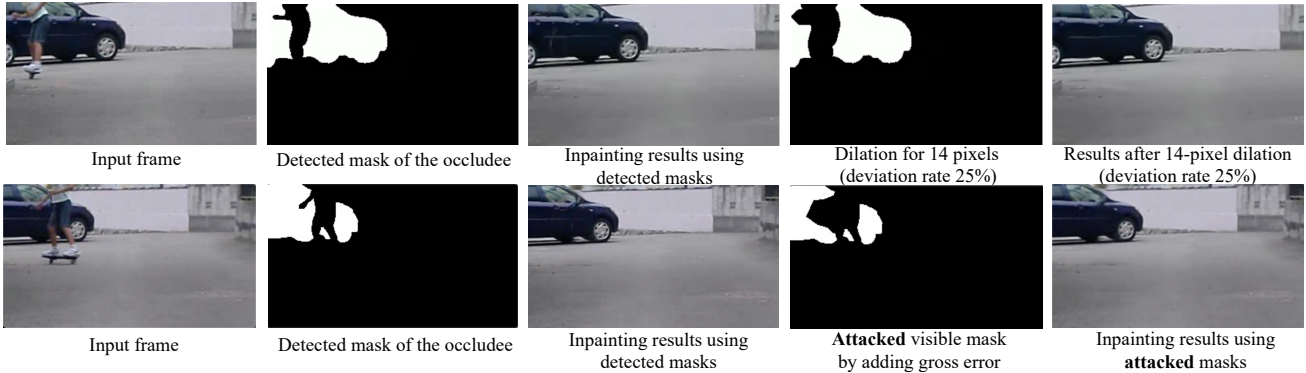| Input frame | Detected mask of the occludee | Inpainting results using detected masks | **Attacked** visible mask by adding gross error | Inpainting results using **attacked** masks |

Figure 9. Sample visual results of VOIN given inaccurate mask segmentation (dilation and gross segmentation errors), which show the robustness of VOIN. Full results are available in the supplementary file.
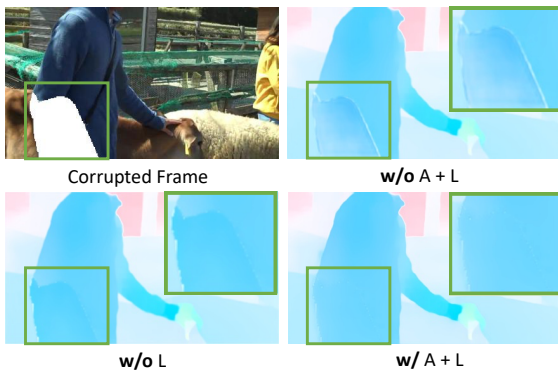

Figure 10. Ablation results of our object flow completion module. A: The predicted amodal object shape guidance. L: Image gradient and warping loss. The corrupted region **w/o** A+L has a large portion of flow error, and the result **w/o** L has unnatural flow transition exhibiting visible seam within the completed foreground.

which in turn improves the inpainting quality during the two-player adversarial process [71].

Effect of Flow-Guided Pixel Propagation. The completed flow warps valid pixels to fill the missing regions in video frames, which greatly reduces the inpainting area and thus the difficulty. The last two rows of Table 4 reflects the accuracy of the completed object flow with large performance gain, where the inpainting result of VOIN under flow guidance w/o adversarial training remains high quality.

5. Conclusion

This paper proposes the new occlusion-aware video object inpainting task with the first large-scale video object
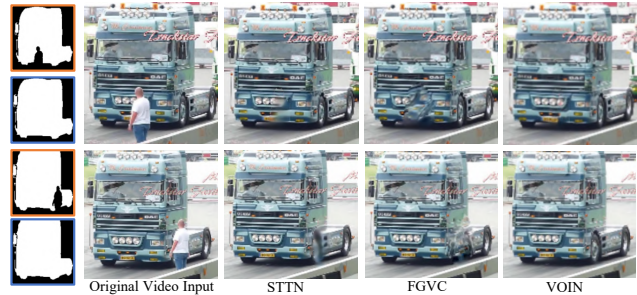

Figure 11. Video scene de-occlusion results comparison, using the shape completion predictions by our VOIN. The left column contains both the visible masks (orange boxes) and predicted complete masks (blue boxes). VOIN detects and recovers the occluded region of the truck with faithful spatial details. The background leaking area (road and shadow under the truck) of the truck occluded by the person's lower body is filled by [3] for fair comparison.

inpainting benchmark *YouTube-VOI* for both training and evaluation. VOIN is a multi-task framework that completes shape and appearance for occluded objects in videos given their visible masks, which contains novel occlusion-aware shape and flow completion modules for propagating temporally-consistent object texture, and a spatio-temporal multi-class discriminator with STAM for enhancing object inpainting quality. We compare VOIN with strong adapted baselines on *YouTube-VOI* benchmark and achieve competitive performance. Our proposed VOIN may benefit many video applications such as video scene de-occlusion/manipulation, and improve video object tracking accuracy under heavy occlusion.

References

[1] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *ICCV*, 2019. 1, 2, 4, 5, 6, 7

[2] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, 2020. 1, 2, 3, 4, 6, 7, 8

[3] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[4] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *CVPR*, 2019. 1, 2, 3, 4, 6, 7

[5] Andréa Aguiar and Renée Baillargeon. Developments in young infants' reasoning about occluded objects. *Cognitive psychology*, 2002. 1

[6] Kanizsa Gaetano. Organization in vision: Essays on gestalt perception. *Praeger Publishers*, 1979. 1

[7] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *CVPR*, 2020. 1, 2, 6, 7

[8] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018. 1, 2

[9] Xiaosheng Yan, Feigege Wang, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *ICCV*, 2019. 1, 2

[10] 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, 4, 6, 7, 8

[11] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *CVPR*, 2017. 1, 2

[12] Patrick Follmann, Rebecca Kö Nig, Philipp Hä Rtinger, Michael Klostermann, and Tobias Bö Tger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *WACV*, 2019. 1

[13] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *CVPR*, 2019. 1, 2

[14] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 1, 6

[15] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time completion of video. *TPAMI*, 29(3):463–476, 2007. 2

[16] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *Siam journal on imaging sciences*, 7(4):1993–2019, 2014. 2

[17] Miguel Granados, James Tompkin, K Kim, Oliver Grau, Jan Kautz, and Christian Theobalt. How not to be seen—object removal from videos of crowded scenes. In *Computer Graphics Forum*, 2012. 2

[18] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[19] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *TPAMI*, 28(7):1150–1163, 2006. 2

[20] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, 2019. 2

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

[22] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *CVPR*, 2019. 2

[23] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *ICCV*, 2019. 2

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

[25] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *CVPR*, 2019. 2, 4, 6

[26] Ke Li and Jitendra Malik. Amodal instance segmentation. In *ECCV*, 2016. 2

[27] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *CVPR*, 2019. 2

[28] Benjamin B Kimia, Ilana Frankel, and Ana-Maria Popescu. Euler spiral for shape completion. *IJCV*, 54(1-3):159–182, 2003. 2

[29] Hongwei Lin, Zihao Wang, Panpan Feng, Xingjiang Lu, and Jinhui Yu. A computational model of topological and geometric recovery for visual curve completion. *Computational Visual Media*, 2(4):329–342, 2016. 2

[30] Nathan Silberman, Lior Shapira, Ran Gal, and Pushmeet Kohli. A contour completion model for augmenting surface reconstructions. In *ECCV*, 2014. 2

[31] Jian Sun, Yin Li, Sing Bing Kang, and Heung-Yeung Shum. Symmetric stereo matching for occlusion handling. In *CVPR*, 2005. 2

[32] John Winn and Jamie Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006. 2

[33] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, 2011. 2

[34] Xianjie Chen and Alan L Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 2

[35] Yi Yang, Sam Hallman, Deva Ramanan, and Charless C Fowlkes. Layered object models for image segmentation. *TPAMI*, 34(9):1731–1743, 2011. 2

[36] Edward Hsiao and Martial Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. *TPAMI*, 36(9):1803–1815, 2014. 2

[37] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang. Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision. In *ECCV*, 2020. 2

[38] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *CVPR*, 2020. 2

[39] Jiyang Qi, Yan Gao, Xiaoyu Liu, Yao Hu, Xinggang Wang, Xiang Bai, Philip HS Torr, Serge Belongie, Alan Yuille, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 2

[40] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV*, 2018. 2

[41] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*, 2018. 2

[42] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, 2015. 2

[43] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *CVPR*, 2020. 2

[44] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, 2021. 2

[45] Tao Yang, Quan Pan, Jing Li, and Stan Z Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *CVPR*, 2005. 2

[46] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012. 2

[47] Tianzhu Zhang, Kui Jia, Changsheng Xu, Yi Ma, and Narendra Ahuja. Partial occlusion handling for visual tracking via robust part matching. In *CVPR*, 2014. 2

[48] Yang Hua, Karteek Alahari, and Cordelia Schmid. Occlusion and motion reasoning for long-term tracking. In *ECCV*, 2014. 2

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2, 3

[50] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019. 3

[51] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *CVPR*, 2020. 3

[52] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 3

[53] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *arXiv preprint arXiv:2011.14503*, 2020. 3

[54] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *ECCV*, 2020. 3

[55] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692*, 2020. 3

[56] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, 2018. 3

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[59] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017. 4

[60] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 4

[61] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 5

[62] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 5

[63] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 2016. 5, 7

[64] Sylvain Paris, Samuel W Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Trans. Graph.*, 30(4):68, 2011. 5

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[66] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 6

[67] John Canny. A computational approach to edge detection. *TPAMI*, (6):679–698, 1986. 6

[68] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *ICCVW*, Oct 2019. 6

[69] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 6, 7

[70] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 6

[71] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 8