Mask-Free Video Instance Segmentation

Lei Ke[1,2] Martin Danelljan[1] Henghui Ding[1] Yu-Wing Tai[2] Chi-Keung Tang[2] Fisher Yu[1]
[1]ETH Zürich [2]HKUST

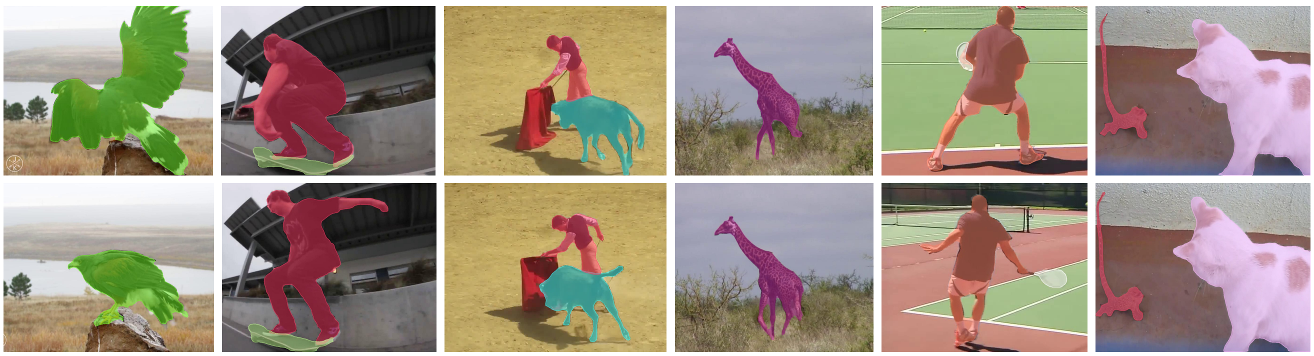

Figure 1. Video instance segmentation (VIS) results of our MaskFreeVIS, trained **without** using any video or image mask annotation. By achieving a remarkable 42.5% mask AP on the YouTube-VIS *val* dataset, with a ResNet-50 backbone, our approach demonstrates that **high-performing VIS can be learned even without any mask annotations**.

Abstract

The recent advancement in Video Instance Segmentation (VIS) has largely been driven by the use of deeper and increasingly data-hungry transformer-based models. However, video masks are tedious and expensive to annotate, limiting the scale and diversity of existing VIS datasets. In this work, we aim to remove the mask-annotation requirement. We propose MaskFreeVIS, achieving highly competitive VIS performance, while only using bounding box annotations for the object state. We leverage the rich temporal mask consistency constraints in videos by introducing the Temporal KNN-patch Loss (TK-Loss), providing strong mask supervision without any labels. Our TK-Loss finds one-to-many matches across frames, through an efficient patch-matching step followed by a K-nearest neighbor selection. A consistency loss is then enforced on the found matches. Our mask-free objective is simple to implement, has no trainable parameters, is computationally efficient, yet outperforms baselines employing, e.g.*, state-of-the-art optical flow to enforce temporal mask consistency. We validate MaskFreeVIS on the YouTube-VIS 2019/2021, OVIS and BDD100K MOTS benchmarks. The results clearly demonstrate the efficacy of our method by drastically narrowing the gap between fully and weakly-supervised VIS performance. Our code and trained models are available at* http://vis.xyz/pub/maskfreevis.

1. Introduction

Video Instance Segmentation (VIS) requires jointly detecting, tracking and segmenting all objects in a video from a given set of categories. To perform this challenging task, state-of-the-art VIS models are trained with complete video annotations from VIS datasets [39, 61, 64]. However, video annotation is costly, in particular regarding object mask labels. Even coarse polygon-based mask annotation is multiple times slower than annotating video bounding boxes [7]. Expensive mask annotation makes existing VIS benchmarks difficult to scale, limiting the number of object categories covered. This is particularly a problem for the recent transformer-based VIS models [6, 17, 57], which tend to be exceptionally data-hungry. We therefore revisit the need for complete mask annotation by studying the problem of weakly supervised VIS *under the mask-free setting*.

While there exist box-supervised instance segmentation models [13, 23, 27, 50], they are designed for images. These weakly-supervised single-image methods do not utilize temporal cues when learning mask prediction, leading to lower accuracy when directly applied to videos. As a source for weakly supervised learning, videos contain much richer information about the scene. In particular, videos adhere to the temporal mask consistency constraint, where the regions corresponding to the same underlying object across different frames should have the same mask label. In this work, we set out to leverage this important constraint for mask-free learning of VIS.

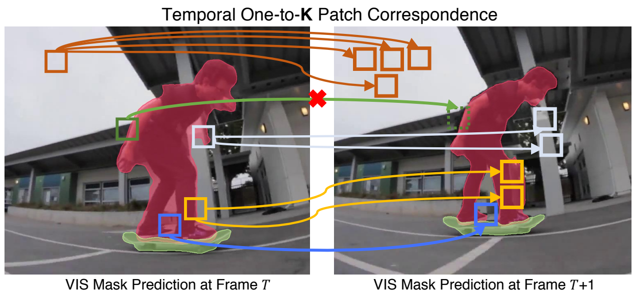

Figure 2. Our Temporal KNN-patch Loss enforces mask consistency between one-to-k patch correspondences found across frames, which allow us to cover the cases where: **(i)** A unique one-to-one match exists (blue); **(ii)** Multiple matches are found due to ambiguities in homogenuous regions (orange) or along image edges (white and yellow); **(iii)** No match is found due to *e.g.* occlusions (green). This allows us to robustly leverage mask consistency constraints in challenging videos.

We propose MaskFreeVIS method, for high performance VIS without any mask annotations. To leverage temporal mask consistency, we introduce the Temporal KNN-patch Loss (TK-Loss), as in Figure 2. To find regions corresponding to the same underlying video object, our TK-Loss first builds correspondences across frames by patch-wise matching. For each target patch, only the top K matches in the neighboring frame with high enough matching score are selected. A temporal consistency loss is then applied to all found matches to promote the mask consistency. Specifically, our surrogate objective function not only promotes the one-to-k matched regions to reach the same mask probabilities, but also commits their mask prediction to a confident foreground or background prediction by entropy minimization. Unlike flow-based models [32,45], which assume one-to-one matching, our approach builds robust and flexible one-to-k correspondences to cope with *e.g.* occlusions and homogeneous regions, *without* introducing additional model parameters or inference cost.

The TK-Loss is easily integrated into existing VIS methods, with no architecture modifications required. During training, our TK-Loss simply replaces the conventional video mask losses in supervising video mask generation. To further enforce temporal consistency through the video clip, TK-Loss is employed in a cyclic manner instead of using dense frame-wise connections. This greatly reduces memory cost with negligible performance drop.

We extensively evaluate our MaskFreeVIS on four large-scale VIS benchmarks, *i.e.*, YouTube-VIS 2019/2021 [61], OVIS [39], and BDD100K MOTS [64]. MaskFreeVIS achieves competitive VIS performance **without** using any video masks or even image mask labels on all datasets. Validated on various methods and backbones, MaskFreeVIS achieves 91.25% performance of its fully supervised counterparts, even outperforming a few recent fully-supervised methods [10, 15, 18, 59] on the popular YTVIS benchmark. Our simple yet effective design greatly narrows the performance gap between weakly-supervised and fully-supervised video instance segmentation. It further demonstrates that expensive video masks, or even image masks, is not necessary for training high-performing VIS models.

Table 1. Mask annotation requirement for state-of-the-art VIS methods. Results are reported using ResNet-50 as backbone on the YTVIS 2019 [61] benchmark. **Video Mask**: using YTVIS video mask labels. **Image Mask**: using COCO [30] image mask labels for image-based pretraining. **Pseudo Video**: using Pseudo Videos from COCO images for joint training [57]. MaskFreeVIS achieves 91.5% (42.5 *vs*. 46.4) of its fully-supervised baseline performance (Mask2Former) **without** using any masks during training.

Method	Video Mask	Image Mask	Pseudo Video	AP
SeqFormer [57]	✓	✓	✓	47.4
VMT [17]	✓	✓	✓	47.9
Mask2Former [6]	✓	✓	✓	47.8
MaskFreeVIS (ours)	✗	✓	✓	46.6
Mask2Former [6]	✓	✓	✗	46.4
MaskFreeVIS (ours)	✗	✗	✗	42.5

Our contributions are summarized as follows: **(i)** To utilize temporal information, we develop a new parameter-free Temporal KNN-patch Loss, which leverages temporal masks consistency using unsupervised one-to-k patch correspondence. We extensively analyze the TK-Loss through ablative experiments. **(ii)** Based on the TK-Loss, we develop the MaskFreeVIS method, enabling training existing state-of-the-art VIS models *without* any mask annotation. **(iii)** To the best of our knowledge, MaskFreeVIS is the first mask-free VIS method attaining high-performing segmentation results. We provide qualitative results in Figure 1. As in Table 1, when integrated into the Mask2Former [6] baseline with ResNet-50, our MaskFreeVIS achieves 42.5% mask AP on the challenging YTVIS 2019 benchmark while using **no** video or image mask annotations. Our approach further scales to larger backbones, achieving 55.3% mask AP on Swin-L backbone with *no* video mask annotations.

We hope our approach will facilitate achieving label-efficient video instance segmentation, enabling building even larger-scale VIS benchmarks with diverse categories by lifting the mask annotation restriction.

2. Related Work

Video Instance Segmentation (VIS) Existing VIS methods can be summarized into three categories: two-stage, one-stage, and transformer-based. Two-stage approaches [2, 18, 28, 29, 61] extend the Mask R-CNN family [11, 19] by designing an additional tracking branch for object association. One-stage works [4, 26, 31, 62] adopt anchor-free detectors [49], generally using linear masks basis combination [3] or conditional mask prediction generation [48]. For the transformer-based models [6, 12, 46, 57, 63], VisTr [54] firstly adapts the transformer [5] for VIS, and IFC [15] further improves its efficiency via memory tokens. Seqformer [57] proposes frame query decomposition while Mask2Former [6] includes masked attention. VMT [17] extends Mask Transfiner [16] to video for high-quality VIS, and IDOL [58] focuses on online VIS. State-of-the-art VIS methods with growing capacity put limited emphasis on weak supervision. In contrast, the proposed MaskFreeVIS is the first method targeting mask-free VIS while attaining competitive performance.

Multiple Object Tracking and Segmentation (MOTS) Most MOTS methods [1, 34, 35, 53, 56] follow the tracking-by-detection principle. PCAN [18] improves temporal segmentation by utilizing space-time memory prototypes, while the one-stage method Unicorn [60] focuses on unification of different tracking frameworks. Compared to the aforementioned fully-supervised MOTS methods, MaskFreeVIS focuses on label efficient training without GT masks by proposing a new surrogate temporal loss which can be easily integrated on them.

Mask-Free VIS Most mask-free instance segmentation works [7, 13, 20, 25, 27, 37, 40, 44] are designed for single images and thus neglect temporal information. Earlier works BoxSup [8] and Box2Seg [22] rely on region proposals produced by MCG [38] or GrabCut [42], leading to slow training. BoxInst [50] proposes the surrogate projection and pixel pairwise losses to replace the original mask learning loss of CondInst [48], while DiscoBox [23] focuses on generating pseudo mask labels guided by a teacher model.

Earlier works have investigated the use of videos for weakly-, semi-, or un-supervised segmentation by leveraging motion or temporal consistency [21, 51, 52]. Most aforementioned approaches do not address the VIS problem, and use optical flow for frame-to-frame matching [24, 32, 43]. In particular, FlowIRN [32] explores VIS using only classification labels and incorporates optical flow to leverage mask consistency. The limited performance makes the class-label only or fully-unsupervised setting difficult to deploy in the real world. SOLO-Track [9] aims to train VIS models without video annotations, and one concurrent work MinVIS [14] performs VIS without video-based model architectures. Unlike the above weakly-supervised training settings, our MaskFreeVIS is designed for eliminating the mask annotation requirement for VIS, as we note that video mask labeling is particularly expensive. MaskFreeVIS enables training VIS models *without* any video masks, or even image masks. Despite its simplicity, MaskFreeVIS drastically reduces the gap between fully-supervised and weakly-supervised VIS models, making weakly-supervised models more accessible in practice.

3. Method

We propose MaskFreeVIS to tackle video instance segmentation (VIS) **without** using any video or even image mask labels. Our approach is generic and can be directly applied to train state-of-the-art VIS methods, such as Mask2Former [6] and SeqFormer [57]. In Sec. 3.1, we first present the core component of MaskFreeVIS: the Temporal KNN-patch Loss (TK-Loss), which leverages temporal consistency to supervise accurate mask prediction, without any human mask annotations. In Sec. 3.2, we then describe how to integrate the TK-Loss with existing spatial weak segmentation losses for transformer-based VIS methods, to achieve mask-free training of VIS approaches. Finally, we introduce image-based pretraining details of our MaskFreeVIS in Sec. 3.2.3.

3.1. MaskFreeVIS

In this section, we introduce the Temporal KNN-patch Loss, illustrated in Figure 3. It serves as an unsupervised objective for mask prediction that leverages the rich spatio-temporal consistency constraints in unlabelled videos.

3.1.1 Temporal Mask Consistency

While an image constitutes a single snapshot of a scene, a video provides multiple snapshots displaced in time. Thereby, a video depicts continuous *change* in the scene. Objects and background move, deform, are occluded, experience variations in lighting, motion blur, and noise, leading to a sequence of different images that are closely related through gradual transformations.

Consider a small region in the scene (Fig. 2), belonging either to an object or background. The pixels corresponding to the projection of this region should have the same mask prediction in every frame, as they belong to the same underlying physical object or background region. However, the aforementioned dynamic changes in the video lead to substantial appearance variations, serving as a natural form of data augmentation. The fact that the pixels corresponding to the same underlying object region should have the same mask prediction under temporal change therefore provides a powerful constraint, *i.e.*, *temporal mask consistency*, which can be used for mask supervision [21, 24, 32, 51, 52].

The difficulty in leveraging the temporal mask consistency constraint stems from the problem of establishing reliable correspondences between video frames. Objects often undergo fast motion, deformations, etc., resulting in substantial appearance change. Furthermore, regions in the scene may be occluded or move out of the image from one frame to the other. In such cases, no correspondence exist. Lastly, videos are often dominated by homogenous regions, such as sky and ground, where the establishment of one-to-one correspondences are error-prone or even ill-defined.

The problem of establishing dense one-to-one correspondences between subsequent video frames, known as optical flow, is a long-standing and popular research topic. However, when attempting to enforce temporal mask consistency through optical flow estimation [24, 32, 43], one encounters two key problems. 1) The one-to-one assumption of optical flow is not suitable in cases of occlusions, homogenous regions, and along single edges, where the corre-

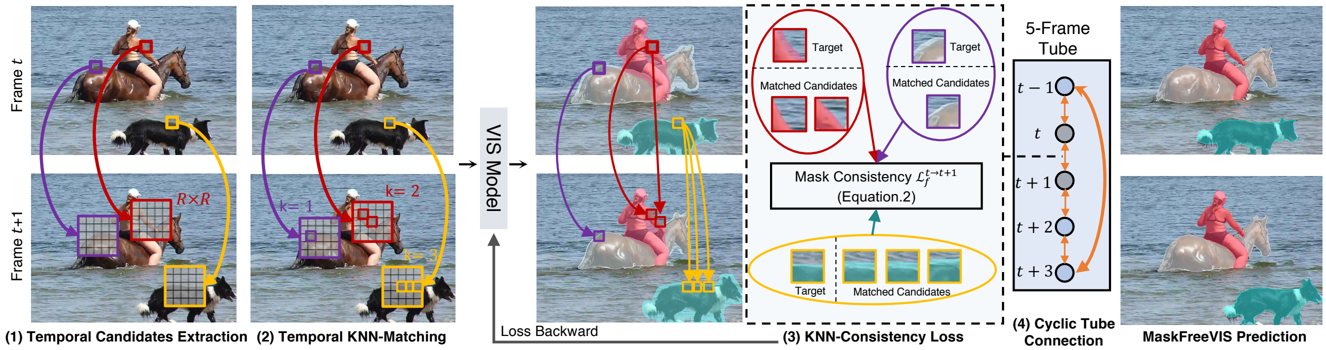

Figure 3. Temporal KNN-patch Loss has four steps: **1)** Patch Candidate Extraction: Patch candidates searching across frames with radius R. **2)** Temporal KNN-Matching: Match k high-confidence candidates by patch affinities. **3)** Consistency loss: Enforce mask consistency objective (Eq. 2) among the matches. **4)** Cyclic Tube Connection: Temporal loss aggregation in the 5-frame tube, detailed in Figure 4.

spondence is either nonexistent, undefined, ambiguous, uncertain, or very difficult to determine. 2) State-of-the-art optical flow estimation rely on large and complex deep networks, with large computational and memory requirements.

Instead of using optical flow, we aim to design a simple, efficient, and parameter-free approach that effectively enforces the temporal mask consistency constraint.

3.1.2 Temporal KNN-patch Loss

Our Temporal KNN-patch Loss (TK-Loss) is based on a simple and flexible correspondence estimation across frames. In contrast to optical flow, we do not restrict our formulation to one-to-one correspondences. Instead, we establish one-to-K correspondences. This include the conventional one-to-one ($K = 1$), where a unique well-defined match exists. However, this allows us to also handle the cases of nonexistent correspondences ($K = 0$) in case of occlusions, and one-to-many ($K \geq 2$) in case of homogenous regions. In cases where multiple matches are found, these most often belong to the same underlying object or background due to their similar appearance, as in Figure 2. This further benefits our mask consistency objective through denser supervision. Lastly, our approach is simple to implement, with negligible computational overhead and no learnable parameters. Our approach is in Figure 3, and contains four main steps, which are detailed next.

1) Patch Candidate Extraction: Let X_p^t denote an $N \times N$ target image patch centered at spatial location $p = (x, y)$ in frame t. Our aim is to find a set of corresponding positions $\mathcal{S}_p^{t\to\hat{t}} = \{\hat{p}_i\}_i$ in frame number $\hat{t}$ that represent the same object region. To this end, we first select candidate locations $\hat{p}$ within a radius R such that $\|p - \hat{p}\| \leq R$. Such windowed patch search exploits spatial proximity across neighboring frames in order to avoid an exhaustive global search. For a fast implementation, the windowed search is performed for all target image patches X_p^t in parallel.

2) Temporal KNN-Matching: We match patch candidate patches through a simple distance computation,

$$\mathbf{d}_{p\to\hat{p}}^{t\to\hat{t}} = \left\| X_p^t - X_{\hat{p}}^{\hat{t}} \right\|, \tag{1}$$

In our ablative experiments (Sec. 4.3), we found the L_2 norm to be the most effective patch matching metric. We select the top K matches with smallest patch distance $\mathbf{d}_{p\to\hat{p}}^{t\to\hat{t}}$. Lastly low-confidence matches are removed by enforcing a maximal patch distance D as $\mathbf{d}_{p\to\hat{p}}^{t\to\hat{t}} < D$. The remaining matches form the set $\mathcal{S}_p^{t\to\hat{t}} = \{\hat{p}_i\}_i$ for each location p.

3) Consistency loss: Let $M_p^t \in [0, 1]$ denote the predicted binary instance mask of an object, evaluated at position p in frame t. To ensure temporal mask consistency constraints, we penalize inconsistent mask predictions between a spatio-temporal point (p, t) and its estimated corresponding points in $\mathcal{S}_p^{t\to\hat{t}}$. In particular we use the following objective,

$$\mathcal{L}_f^{t\to\hat{t}} = \frac{1}{HW} \sum_p \sum_{\hat{p}_i \in \mathcal{S}_p^{t\to\hat{t}}} L_{\text{cons}}(M_p^t, M_{\hat{p}_i}^{\hat{t}}), \tag{2}$$

where mask consistency is measured as

$$L_{\text{cons}}(M_p^t, M_{\hat{p}}^{\hat{t}}) = -\log\left(M_p^t M_{\hat{p}}^{\hat{t}} + (1 - M_p^t)(1 - M_{\hat{p}}^{\hat{t}})\right). \tag{3}$$

Note that Eq. (3) only attains its minimum value of zero if both predictions indicate exactly background ($M_p^t = M_{\hat{p}}^{\hat{t}} = 0$) or foreground ($M_p^t = M_{\hat{p}}^{\hat{t}} = 1$). The objective does thus not only promote the two mask predictions to achieve the same probability value $M_p^t = M_{\hat{p}}^{\hat{t}}$, but also to commit to a certain foreground or background prediction.

4) Cyclic Tube Connection: Suppose the temporal tube consists of T frames. We compute the temporal loss for the whole tube in a cyclic manner, as in Figure 4. The start frame is connected to the end frame, which introduces direct long-term mask consistency across the two temporally most distant frames. The temporal TK-Loss for the whole tube is given by

$$\mathcal{L}_{\text{temp}} = \sum_{t=1}^{T} \begin{cases} \mathcal{L}_f^{t\to(t+1)} & \text{if } t < T - 1 \\ \mathcal{L}_f^{t\to 0} & \text{if } t = T - 1. \end{cases} \tag{4}$$

Compared to inter-frame dense connections in Figure 4, we find the cyclic loss to achieve similar performance but greatly reduce the memory usage as validated in the experiment section.

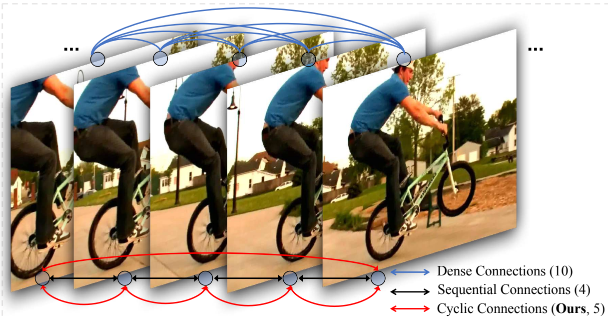

Figure 4. Illustration of different frame-wise tube connection settings (connection number) in the temporal loss design.

3.2. Training MaskFreeVIS

In this section, we describe how to train state-of-the-art VIS methods using our TK-Loss, **without** any mask annotations. Our MaskFreeVIS approach is jointly supervised with spatial-temporal surrogate losses, and is easily integrated with existing transformer-based methods. We also detail mask-free image-based pre-training for MaskFreeVIS to fully eliminate mask usage during training.

3.2.1 Joint Spatio-temporal Regularization

To train MaskFreeVIS, in addition to our proposed Temporal KNN-patch Loss for temporal mask consistency, we leverage existing spatial weak segmentation losses to jointly enforce intra-frame consistency.

Spatial Consistency To explore spatial weak supervision signals from image bounding boxes and pixel color, we utilize the representative Box Projection Loss L_{proj} and Pairwise Loss L_{pair} in [50], to replace the supervised mask learning loss. The Projection Loss L_{proj} enforces the projection P' of the object mask onto the $\vec{x}$-axis and $\vec{y}$-axis of image to be consistent with its ground-truth box mask. For the temporal tube with T frames, we concurrently optimize all predicted frame masks of the tube as,

$$\mathcal{L}_{\text{proj}} = \sum_{t=1}^{T} \sum_{d \in \{\vec{x}, \vec{y}\}} D(P'_d(M_p^t), P'_d(M_b^t)), \tag{5}$$

where D denotes dice loss, P' is the projection function along $\vec{x}/\vec{y}$-axis direction, M_p^t and M_b^t denote predicted instance mask and its GT box mask at frame t respectively. The object instance index is omitted here for clarity.

The Pairwise Loss L_{pair}, on the other hand, constrains spatially neighboring pixels of single frame. For pixel of locations p'_i and p'_j of with color similarity $\geqslant \sigma_{\text{pixel}}$, we enforce their predicted mask labels to be consistent, following Eq. (3) as,

$$\mathcal{L}_{\text{pair}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{p'_i \in H \times W} L_{\text{cons}}(M_{p'_i}^t, M_{p'_j}^t). \tag{6}$$

The spatial losses are combined with a weight factor λ_{pair}:

$$\mathcal{L}_{\text{spatial}} = \mathcal{L}_{\text{proj}} + \lambda_{\text{pair}} \mathcal{L}_{\text{pair}}. \tag{7}$$

Temporal Consistency We adopt the Temporal KNN-patch Loss in Sec. 3.1.2 as $\mathcal{L}_{\text{temp}}$ to leverage temporal mask consistency. The overall spatio-temporal objective $\mathcal{L}_{\text{seg}}$ for optimizing video segmentation is summarized as,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{spatial}} + \lambda_{\text{temp}} \mathcal{L}_{\text{temp}}. \tag{8}$$

3.2.2 Integration with Transformer-based Methods

Existing works [13, 48] on box-supervised segmentation losses are coupled with either one-stage or two-stage detectors, such as Faster R-CNN [41] and CondInst [48], and only address the single image case. However, state-of-the-art VIS methods [6, 57] are based on transformers. These works perform object detection via set prediction, where predicted instance masks need to be matched with mask annotations when evaluating the loss. To integrate mask-free VIS training with transformers, one key modification is in this instance-sequence matching step.

Since only ground-truth bounding boxes are available for box sequence matching, as an initial attempt, we first produce bounding box predictions from the estimated instance masks. Then, we employ the sequential box matching cost function used in VIS methods [55, 57]. To compute matching cost for whole sequence, $\mathcal{L}_1$ loss and generalized IoU loss for each individual bounding box is averaged across the frames. However, we observe the matching results of frame-wise averaging can easily be affected by a single outlier frame, especially under weak segmentation setup, leading to instability during training and performance decrease.

Spatio-temporal Box Mask Matching Instead of using the aforementioned frame-wise matching, we empirically find spatio-temporal box-to-mask matching to produce substantial improvement under the weak segmentation setting. We first convert each predicted instance mask to a bounding box mask, and convert the ground-truth box to box mask. We then randomly sample a equal number of points from the ground-truth box mask sequence and predicted box mask sequence, respectively. Different from Mask2Former [6], we only adopt the dice IoU loss to compute sequence matching cost. We find that cross-entropy accumulates errors per pixel, leading to imbalanced values between large and small objects. In contrast, the IoU loss in normalized per object, leading to a balanced metric. We study different instance sequence matching strategies under the mask-free VIS setting in the ablation experiments.

3.2.3 Image-based MaskFreeVIS Pre-training

Most VIS models [6, 57, 61] are initialized from a model pretrained on the COCO instance segmentation dataset. To completely eliminate mask supervision, we pretrain our MaskFreeVIS on COCO using only box supervision. We adopt the spatial consistency loss described in Sec. 3.2.1 on single frame to replace the original GT mask losses in Mask2Former [6], while following the same image-based training setup on COCO. Thus, we provide two training settings in our experiments, one eliminates both image and video mask during training, while the other adopts weights pretrained with COCO mask annotations. In both cases, no video mask annotations are used.

4. Experiments

4.1. Datasets

YTVIS 2019/2021 We perform experiments on the large-scale YouTube-VIS [61] 2019 and 2021. YTVIS 2019 includes 2,883 videos of 131k annotated object instances belonging to 40 categories. To handle more complex cases, YTVIS 2021 updates YTVIS 2019 with additional 794 videos for training and 129 videos for validation, including more tracklets with confusing motion trajectories.
OVIS We also train and evaluate on OVIS [39], a VIS benchmark on occlusion learning. OVIS consists of instance masks covering 25 categories with 607, 140 and 154 videos for train, valid and test respectively.
BDD100K MOTS We further report results of MaskFreeVIS on the large-scale self-driving benchmark BDD100K [64] MOTS. The dataset annotates 154 videos (30,817 images) for training, 32 videos (6,475 images) for validation, and 37 videos (7,484 images) for testing.

4.2. Implementation Details

Our proposed approach only requires replacing the original video mask loss in state-of-the-art VIS methods. In particular, we adopt Mask2Former [6] and SeqFormer [57] due to their excellent VIS results. Unless specified, we kept all other training schedules and settings the same as in the original methods. For the Temporal KNN-patch Loss, we set the patch size to 3×3, search radius to 5 and $K = 5$. We adopt the L_2 distance as the patch matching metric and set the matching threshold to 0.05. On YTVIS 2019/2021, the Mask2Former based models are trained with AdamW [33] with learning rate 10^{-4} and weight decay 0.05. The learning rate decays by 10 times at with a factor of 2/3. We set batch size to 16, and train 6k/8k iterations on YTVIS 2019/2021. For experiments on OVIS and BDD100K, we adopted the COCO mask pretrained models by VITA and Unicorn. For the sampled temporal tube at training, we use 5 frames with shuffling instead of 2 frames for better temporal regularization. The compared baselines are adjusted accordingly. During testing, since there is no architecture modification, the inference of MaskFreeVIS is the same to the baselines. More details are in the Supp. file.

4.3. Ablation Experiments

We perform detailed ablation studies for MaskFreeVIS using ResNet-50 as backbone on the YTVIS 2019 *val* set. We adopt the COCO box-pretrained model as initialization to eliminate all mask annotations from the training. Taking Mask2Former [6] as the base VIS method, we analyze the impact of individual proposed components. Moreover, we study several alternative solutions for temporal matching and influence of different hyper-parameters to TK-Loss.

Table 2. Different temporal matching schemes under the mask-free training setting on YTVIS2019 val. 'Param' indicates whether the matching scheme brings extra model parameters.

Temporal Matching Scheme	Param	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Baseline (**No** Matching)	✗	38.6	65.9	38.8	38.4	47.7
Flow-based Matching	✓	40.2	66.3	41.9	40.5	49.1
Temporal Deformable Matching	✓	39.6	65.9	40.1	39.9	48.6
Learnable Pixel Matching	✓	39.5	65.7	40.2	39.7	48.4
Learnable Patch Matching	✓	40.6	66.5	42.6	40.3	49.2
3D Pairwise Matching	✗	39.4	65.0	41.7	40.2	48.0
Temporal KNN-Matching (Ours)	✗	**42.5**	**66.8**	**45.7**	**41.2**	**51.2**

Comparison on Temporal Matching Schemes Table 2 compares our Temporal KNN-Matching to four alternative frame-wise matching approaches for enforcing temporal mask consistency. **Flow-based Matching** employs the pretrained optical flow model RAFT [45] to build pixel correspondence [32]. **Temporal Deformable Matching** adopts the temporal deformable kernels [47] to predict the pixel offsets between the target and alignment frame. Instead of using raw patches, **Learnable Pixel/Patch Matching** employs jointly learned deep pixel/patch embeddings via three learnable FC layers, which are then used to compute the affinities in a soft attention-like manner. **3D Pairwise Matching** directly extends $\mathcal{L}_{\text{pair}}$ designed for spatial images to the temporal dimension, where pairwise affinity loss is computed among pixels not only in within the frame but also across multiple frames.

In Table 2, compared to Flow-based Matching with one-to-one pixel correspondence, our paramter-free Temporal KNN-Matching with one-to-K improves by about 2.3 AP. The prediction of flow-based models are not reliable in case of occlusions and homogeneous regions, and are also influenced by the gap between the training dataset and real-world video data. For the above mentioned deformable and learnable matching schemes, since there are only weak bounding box labels during training, the temporal matching relation is learnt implicitly. We empirically observe the instability during training with limited improvement under the mask-free training setting. For direct generalization of $\mathcal{L}_{\text{pair}}$, it only leads to 0.8 mask AP performance improvement. Despite the simplicity and efficiency of the TK-loss, it significantly improves VIS performance by 3.9 mask AP.
Effect of Temporal KNN-patch Loss MaskFreeVIS is trained with joint spatio-temporal losses. In Table 3, to evaluate the effectiveness of each loss component, we compare the performance of MaskFreeVIS solely under the spatial pairwise loss [50] or our proposed TK-Loss. Compared to the 2.0 mask AP improvement by the spatial pairwise loss, the TK-Loss substantially promotes the mask AP from 36.6 to 41.6, showing the advantage of our flexible one-to-K patch correspondence design in leveraging temporal consistency. We show the VIS results in Figure 5 to visualize

Table 3. Effect of the Spatial Pairwise loss and our Temporal KNN-patch Loss on YTVIS2019 val.

Box Proj	Pairwise	TK-Loss	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
✓			36.6	66.5	36.2	37.1	45.0
✓	✓		$38.6_{\uparrow 2.0}$	65.9	38.8	38.4	47.7
✓		✓	$41.6_{\uparrow \mathbf{5.0}}$	**68.4**	43.5	40.1	50.5
✓	✓	✓	**42.5**$_{\uparrow 5.9}$	66.8	**45.7**	**41.2**	**51.2**

Table 4. Effect of using max **K** matches on YTVIS2019 val.

K	AP	AP_{50}	AP_{75}	AR_1
1	40.8	65.8	44.1	40.3
3	41.9	66.9	45.1	**41.9**
5	**42.5**	66.8	**45.7**	41.2
7	42.3	**67.1**	44.6	40.6

Table 5. Patch Matching metrics comparison. NCC is Norm. Cross-Correlation.

Metric	AP	AP_{50}	AP_{75}	AR_1
NCC	41.7	66.7	43.4	**41.4**
L_1	41.2	66.2	43.6	40.3
L_2	**42.5**	**66.8**	**45.7**	41.2

Table 6. Impact of search radius **R** on YTVIS2019 val.

R	AP	AP_{50}	AP_{75}	AR_1
1	39.6	65.7	40.2	39.7
3	41.3	66.6	43.0	40.3
5	**42.5**	66.8	**45.7**	**41.2**
7	42.3	**67.1**	44.6	40.6

Table 7. Influence of patch size **N** on YTVIS2019 val.

N	AP	AP_{50}	AP_{75}	AR_1
1	40.1	65.2	42.2	40.0
3	**42.5**	66.8	**45.7**	41.2
5	42.1	**68.7**	44.4	**42.1**
7	41.5	66.3	44.8	41.3

the effectiveness of each loss component.

Analysis of Temporal KNN-patch Loss In Table 4, we study the influence of **K**, the maximum number of matches selected in Temporal KNN-patch Loss. The best result is obtained for $K = 5$, while $K = 1$ only allows for the one-to-one and no-match cases. The improvement from 40.8 mask AP to 42.5 mask AP reveals the benefit brought by the flexible one-to-many correspondence design. We also analyze the matching metric in Table 5, search radius in Table 6, and patch size in Table 7 (see Sec. 3.1.2 for details). When patch size N is increased from 1 to 3 in Table 7, the performance of MaskFreeVIS is improved by 2.4 AP, validating the importance of patch structure in robust matching.

Effect of the Cyclic Tube Connection We compare three frame-wise tube connection schemes (Figure 4) for the TK-Loss in Table 8. While dense connection brings forth the best performance, it doubles the training memory with minor improvement compared to Cyclic connection. Comparing to Sequential connection, our Cyclic connection benefits from long-range consistency, improving the performance of 0.6 mask AP with an affordable memory cost growth.

Table 8. Comparison of the tube connection schemes, illustrated in Fig. 4. The tube consist of 5 frames. 'Mem' denotes the memory consumption per sampled video by the TK-Loss during training.

Tube Connect	Connect Num.	Mem (MB)	AP	AP_{50}	AP_{75}	AR_1
Dense	10	1526	42.7	68.0	44.3	41.5
Sequential	4	631	41.9	66.5	44.7	41.2
Cyclic (Ours)	5	773	42.5	66.8	45.7	41.2

Comparison on Sequence Matching Functions Besides TK-Loss, we analyze the influence of sequence matching cost functions for transformer-based VIS methods under the mask-free setting. In Table 9, we identify the substantial advantage of Spatio-temporal box-mask matching over frame-wise cost averaging [55, 57]. As discussed in Sec. 3.2.2, we

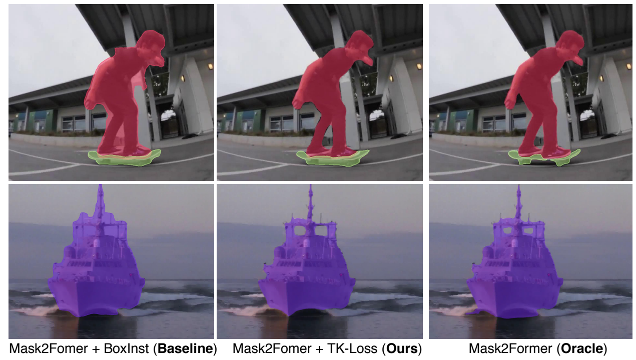

Figure 5. Qualitative results comparison between using Spatial Pairwise loss [50], our TK-Loss, and Mask2Former (oracle) trained with GT video and image masks.

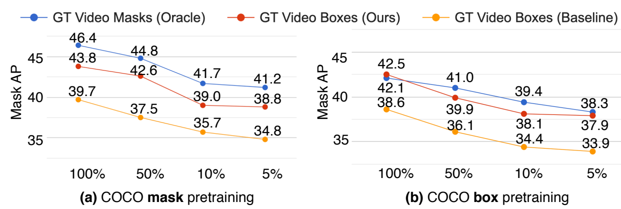

Figure 6. Results on YTVIS 2019 val with various percentages of the YTVIS training data. Baseline denotes Mask2Former [6] trained with GT video boxes using BoxInst [50], while Oracle denotes Mask2Former trained with GT video masks.

Table 9. Comparison of Set Matching Cost Functions on YTVIS2019 val. ST-BoxMask denotes our Spatio-temporal Box Mask matching. **w/o** CE denotes removing cross-entropy cost.

Matching Cost Function	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Frame-wise Averaging [55, 57]	37.6	64.2	39.5	37.5	45.7
ST-BoxMask	40.8	**67.8**	42.2	40.0	49.2
ST-BoxMask **w/o** CE	**42.5**	66.8	**45.7**	**41.2**	**51.2**

achieve further gain by removing the object size imbalanced cross-entropy cost computation.

Training on Various Amounts of Data To further study label-efficient VIS, we validate the effect of MaskFreeVIS under various percentages of the YTVIS 2019 training data. We uniformly sample frames and their labels for each video, and set the minimum sampled number of frames to 1. Figure 6 presents the experimental results, which shows the consistent large improvement (over 3.0 AP) brought by our TK-Loss under various amount of data. In particular, we note that our approach with 50% data even outperforms the fully-supervised model with 10% training data.

4.4. Comparison with State-of-the-art Methods

We compare MaskFreeVIS with the state-of-the-art fully/weakly supervised methods on benchmarks YTVIS 2019/2021, OVIS and BDD100K MOTS. We integrate MaskFreeVIS on four representative methods [6, 12, 57, 60], attaining consistent large gains over the strong baselines.

YTVIS 2019/2021 Table 10 compares the performance on YTVIS 2019. Using R50/R101 as backbone and with the same training setting, MaskFreeVIS achieves 42.5/45.8 AP, improving 3.9/5.0 AP over the strong baseline adopt-

Table 10. Comparison on YTVIS 2019. I: using COCO mask pretrained model as initialization. V: using YTVIS video masks during training. $*$: using pseudo mask from COCO images for joint training [57]. M2F: Mask2Former [6], SeqF: SeqFormer [57].

Method	Mask ann.	Back-bone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Fully-supervised:							
PCAN [18]	I+V	R50	36.1	54.9	39.4	36.3	41.6
EfficientVIS [59]	I+V	R50	37.9	59.7	43.0	40.3	46.6
InsPro [10]	I+V	R50	40.2	62.9	43.1	37.6	44.5
IFC [15]	I+V	R50	42.8	65.8	46.8	43.8	51.2
VMT* [17]	I+V	R50	47.9	-	52.0	45.8	-
SeqF* [57]	I+V	R50	47.4	69.8	51.8	45.5	54.8
M2F	I+V	R50	46.4	68.0	50.0	-	-
M2F*	I+V	R50	47.8	69.2	52.7	46.2	56.6
Prev. Weakly-supervised:							
FlowIRN [32]	-	R50	10.5	27.2	6.2	12.3	13.6
SOLO-Track [9]	I	R50	30.6	50.7	33.5	31.6	37.1
Mask-free:							
M2F + Flow Consist [45]	-	R50	40.2	66.3	41.9	40.5	49.1
M2F + BoxInst [50]	-	R50	38.6	64.2	38.5	38.0	46.8
M2F + **MaskFreeVIS**	-	R50	$42.5_{\uparrow 3.9}$	66.8	45.7	41.2	51.2
M2F + **MaskFreeVIS**	I	R50	$43.8_{\uparrow 5.2}$	70.7	46.9	41.5	52.3
M2F + **MaskFreeVIS***	I	R50	$\mathbf{46.6}_{\uparrow 8.0}$	**72.5**	**49.7**	**44.9**	**55.7**
Fully-supervised:							
M2F	V	R101	45.6	72.6	48.9	44.3	54.5
M2F	I+V	R101	49.2	72.8	54.2	-	-
M2F*	I+V	R101	49.8	73.6	55.4	48.0	58.0
SeqF*	I+V	R101	49.0	71.1	55.7	46.8	56.9
Mask-free:							
M2F + BoxInst [50]	-	R101	40.8	67.8	42.2	40.0	49.2
M2F + **MaskFreeVIS**	-	R101	$45.8_{\uparrow 5.0}$	70.8	48.6	**45.3**	55.2
M2F + **MaskFreeVIS**	I	R101	$47.3_{\uparrow 6.5}$	**75.4**	49.9	44.6	55.2
M2F + **MaskFreeVIS***	I	R101	$\mathbf{48.9}_{\uparrow 8.1}$	74.9	**54.7**	44.9	57.0
SeqF + **MaskFreeVIS***	I	R101	48.6	74.0	52.2	45.9	**57.2**
Fully-supervised:							
M2F	I+V	SwinL	60.4	84.4	67.0	-	-
Mask-free:							
M2F + BoxInst [50]	-	SwinL	49.8	73.2	55.5	48.2	58.1
M2F + **MaskFreeVIS**	-	SwinL	$54.3_{\uparrow 4.5}$	82.6	61.1	50.2	61.3
M2F + **MaskFreeVIS***	I	SwinL	$\mathbf{55.3}_{\uparrow 5.5}$	82.5	60.8	50.7	62.2

ing BoxInst [50] losses. MaskFreeVIS, **without** any mask labels, even *significantly outperforms some recent fully-supervised methods* such as EfficientVIS [59] and InsPro [10]. On R50/R101/Swin-L, our MaskFreeVIS consistently attains over 91% of its fully-supervised counterpart trained with both GT image and video masks. We also observe similar larger performance growth over the baseline on YTVIS 2021 in Table 11. The excellent performance substantially narrows the performance gap between fully-supervised and weakly-supervised VIS.

OVIS We also conduct experiments on OVIS in Table 12 using R50 as backbone. We integrate MaskFreeVIS with VITA [12], promoting the baseline performance from 12.1 to 15.7 under the mask-free training setting.

BDD100K MOTS Table 13 further validates our approach on BDD100K MOTS. Integrated with Unicorn [60], MaskFreeVIS achieves 23.8 mMOTSA by improving over 4.9 points over the strong baseline and thus surpassing the fully-supervised QDTrack-mots [36]. The consistent large gains on four benchmarks and four base VIS methods validates the generalizability of our MaskFreeVIS.

[1]This work is supported in part by Research Grant Council of the Hong Kong SAR under grant no. 16201420 and Huawei Technologies.

Table 11. Comparison on YTVIS 2021. Refer to Table 10 for the symbol abbreviations.

Method	Mask ann.	Back-bone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Fully-supervised:							
MaskTrack [61]	I+V	R50	28.6	48.9	29.6	26.5	33.8
IFC [15]	I+V	R50	36.6	57.9	39.3	-	-
SeqF* [57]	I+V	R50	40.5	62.4	43.7	36.1	48.1
M2F	I+V	R50	40.6	60.9	41.8	-	-
Mask-free:							
M2F + BoxInst [50]	-	R50	32.1	52.8	34.4	31.0	38.1
M2F + **MaskFreeVIS**	-	R50	$36.2_{\uparrow 4.1}$	60.8	39.2	34.6	45.6
M2F + **MaskFreeVIS**	I	R50	$37.2_{\uparrow 5.1}$	61.9	40.3	35.3	46.1
M2F + **MaskFreeVIS***	I	R50	$\mathbf{40.9}_{\uparrow 8.8}$	**65.8**	**43.3**	**37.1**	**50.5**
Fully-supervised:							
M2F	I+V	R101	42.4	65.9	45.8	-	-
Mask-free:							
M2F + BoxInst [50]	-	R101	33.3	55.2	32.5	32.1	41.9
M2F + **MaskFreeVIS**	-	R101	$37.3_{\uparrow 4.0}$	61.6	39.4	34.1	45.6
M2F + **MaskFreeVIS**	I	R101	$38.2_{\uparrow 4.9}$	62.4	40.0	34.9	46.2
M2F + **MaskFreeVIS***	I	R101	$\mathbf{41.6}_{\uparrow 8.3}$	**66.2**	**44.8**	**36.3**	**49.2**

Table 12. State-of-the-art comparison on the OVIS using R50.

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Fully-supervised:					
CrossVIS [62]	14.9	32.7	12.1	10.3	19.8
Mask2Former [6]	17.3	37.3	15.1	10.5	23.5
VMT [17]	16.9	36.4	13.7	10.4	22.7
VITA [12]	19.6	41.2	17.4	11.7	26.0
Video Mask-free:					
VITA [12] + BoxInst [50]	12.1	28.3	10.2	8.8	17.9
VITA [12] + **MaskFreeVIS**	$\mathbf{15.7}_{\uparrow 3.6}$	**35.1**	**13.1**	**10.1**	**20.4**

Table 13. State-of-the-art comparison on the BDD100K segmentation tracking validation set.

Method	mMOTSA↑	mMOTSP↑	mIDF↑	ID sw.↓	mAP↑
Fully-supervised:					
STEm-Seg [1]	12.2	58.2	25.4	8732	21.8
QDTrack-mots-fix [36]	23.5	66.3	44.5	973	25.5
PCAN [18]	27.4	66.7	45.1	876	26.6
Unicorn [60]	29.6	67.7	44.2	1731	32.1
Video Mask-free:					
Unicorn [60] + BoxInst [50]	18.9	58.7	36.3	3298	22.1
Unicorn [60] + **MaskFreeVIS**	$\mathbf{23.8}_{\uparrow 4.9}$	**66.7**	**44.9**	**2086**	**24.8**

5. Conclusion

MaskFreeVIS is the first competitive VIS method that does not need *any* mask annotations during training. The strong results lead to a remarkable conclusion: mask labels are not a necessity for high-performing VIS. Our key component is the unsupervised Temporal KNN-patch Loss, which replaces the conventional video masks losses by leveraging temporal mask consistency constraints. Our approach greatly reduces the long-standing gap between fully-supervised and weakly-supervised VIS on four large-scale benchmarks. MaskFreeVIS thus opens up many opportunities for researchers and practitioners for label-efficient VIS.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 3, 8

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 2

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 2

[4] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 2020. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[6] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1, 2, 3, 5, 6, 7, 8

[7] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *CVPR*, 2022. 1, 3

[8] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 3

[9] Yang Fu, Sifei Liu, Umar Iqbal, Shalini De Mello, Humphrey Shi, and Jan Kautz. Learning to track instances without video annotations. In *CVPR*, 2021. 3, 8

[10] Fei He, Haoyang Zhang, Naiyu Gao, Jian Jia, Yanhu Shan, Xin Zhao, and Kaiqi Huang. Inspro: Propagating instance query and proposal for online video instance segmentation. In *NeurIPS*, 2022. 2, 8

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[12] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. In *NeurIPS*, 2022. 2, 7, 8

[13] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *NeurIPS*, 2019. 1, 3, 5

[14] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 3

[15] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In *NeurIPS*, 2021. 2, 8

[16] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *CVPR*, 2022. 3

[17] Lei Ke, Henghui Ding, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Video mask transfiner for high-quality video instance segmentation. In *ECCV*, 2022. 1, 2, 3, 8

[18] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. In *NeurIPS*, 2021. 2, 3, 8

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

[20] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 3

[21] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *ICLR*, 2022. 3

[22] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *ECCV*, 2020. 3

[23] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *ICCV*, 2021. 1, 3

[24] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In *ICCV*, 2019. 3

[25] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *CVPR*, 2021. 3

[26] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *CVPR*, 2021. 2

[27] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. Box-supervised instance segmentation with level set evolution. In *ECCV*, 2022. 1, 3

[28] Chung-Ching Lin, Ying Hung, Rogerio Feris, and Linglin He. Video instance segmentation tracking with a modified vae architecture. In *CVPR*, 2020. 2

[29] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *CVPR*, 2021. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[31] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*, 2021. 2

[32] Qing Liu, Vignesh Ramanathan, Dhruv Mahajan, Alan Yuille, and Zhenheng Yang. Weakly supervised instance segmentation for videos with temporal mask consistency. In *CVPR*, 2021. 2, 3, 6, 8

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[34] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022. 3

[35] Anton Milan, Laura Leal-Taixé, Konrad Schindler, and Ian Reid. Joint tracking and segmentation of multiple targets. In *CVPR*, 2015. 3

[36] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021. 8

[37] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015. 3

[38] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE TPAMI*, 39(1):128–140, 2016. 3

[39] 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*, 2021. 1, 2, 6

[40] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2016. 3

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5

[42] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23(3):309–314, 2004. 3

[43] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In *ICCV*, 2017. 3

[44] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *CVPR*, 2019. 3

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2, 6, 8

[46] Omkar Thawakar, Sanath Narayan, Jiale Cao, Hisham Cholakkal, Rao Muhammad Anwer, Muhammad Haris Khan, Salman Khan, Michael Felsberg, and Fahad Shahbaz Khan. Video instance segmentation via multi-scale spatio-temporal split attention transformer. In *ECCV*, 2022. 2

[47] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 6

[48] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 2, 3, 5

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 2

[50] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *CVPR*, 2021. 1, 3, 5, 6, 7, 8

[51] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning semantic segmentation with weakly-annotated videos. In *ECCV*, 2016. 3

[52] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *ECCV*, 2016. 3

[53] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019. 3

[54] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2

[55] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 5, 7

[56] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *CVPR*, 2021. 3

[57] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7, 8

[58] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 3

[59] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. In *CVPR*, 2022. 2, 8

[60] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022. 3, 7, 8

[61] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 5, 6, 8

[62] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *ICCV*, 2021. 2, 8

[63] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *CVPR*, 2022. 2

[64] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 1, 2, 6