Integrating Boxes and Masks: A Multi-Object Framework for Unified Visual Tracking and Segmentation

Yuanyou Xu[1,2†], Zongxin Yang[1], Yi Yang[1‡]
[1]ReLER, CCAI, Zhejiang University, China [2]Baidu Research, China
{yoxu, zongxinyang, yangyics}@zju.edu.cn

Abstract

Tracking any given object(s) spatially and temporally is a common purpose in Visual Object Tracking (VOT) and Video Object Segmentation (VOS). Joint tracking and segmentation have been attempted in some studies but they often lack full compatibility of both box and mask in initialization and prediction, and mainly focus on single-object scenarios. To address these limitations, this paper proposes a Multi-object Mask-box Integrated framework for unified Tracking and Segmentation, dubbed MITS. Firstly, the unified identification module is proposed to support both box and mask reference for initialization, where detailed object information is inferred from boxes or directly retained from masks. Additionally, a novel pinpoint box predictor is proposed for accurate multi-object box prediction, facilitating target-oriented representation learning. All target objects are processed simultaneously from encoding to propagation and decoding, as a unified pipeline for VOT and VOS. Experimental results show MITS achieves state-of-the-art performance on both VOT and VOS benchmarks. Notably, MITS surpasses the best prior VOT competitor by around 6% on the GOT-10k test set, and significantly improves the performance of box initialization on VOS benchmarks. The code is available at https://github.com/yoxu515/MITS.

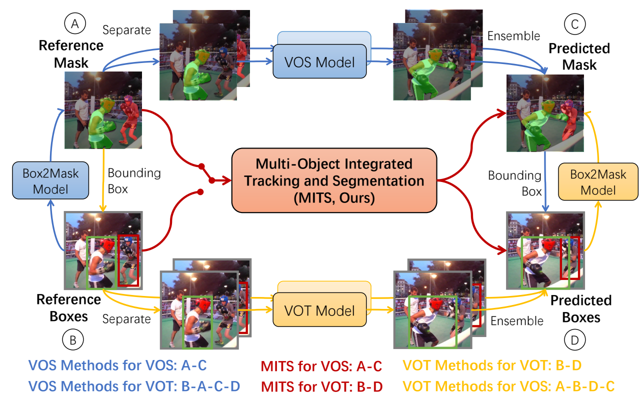

Figure 1. We propose a Multi-object Mask-box Integrated framework for unified Tracking and Segmentation (MITS). Compared with VOT or VOS methods, our framework *integrates both boxes and masks* for initialization and prediction in a *unified manner*, without any extra model for box-to-mask conversion, and supports processing *multiple objects* simultaneously for VOT and VOS.

1. Introduction

Visual object tracking (VOT) [57, 23, 32, 33] and video object segmentation (VOS) [63, 84, 79, 37] are two critical tasks in computer vision. Visual object tracking involves identifying and tracking specific object(s) in a video stream over time. Video object segmentation aims to segment given object(s) in a video sequence and separate it from the background. Both tasks are essential for applications such as video surveillance and autonomous driving.

In VOT, object sizes and positions are indicated with boxes as box representation, while in VOS object shapes and contours are marked with pixel-level masks as mask representation. Despite their differences, VOT and VOS share similarities. Both tasks require the ability to identify and locate the target object in a video stream accurately in spatial dimension, and to be robust against challenges such as occlusion and fast motion in temporal dimension.

In view of the similarity, tracking and segmentation may have unified multi-knowledge representations [90] and have been explored jointly in some works. 1) *Unification*. A straightforward solution is to perform conversion between boxes and masks to utilize VOT methods on VOS or VOS methods on VOT. A mask can be converted to a box easily, but generating a mask from a box is hard. Some methods were proposed to address the box-to-mask estimation problem [47, 88, 101]. However, separate but not unified models hinder end-to-end training and are inconvenient to manage in practical applications. 2) *Compatibility*. Several studies have attempted to unify these two tasks into a single frame-

[†]Yuanyou Xu worked on this at his Baidu Research internship.
[‡]Yi Yang is the corresponding author.

work. However, some of them [78, 75, 62] still lack compatibility and flexibility in box/mask input/output and resort to extra models. 3) *Multi-Object*. Despite that some methods [48, 81, 86] possess strong compatibility across VOT and VOS, they mainly focus on the single object scenario and use an ensemble strategy to aggregate the separate result of each object in the multiple object scenario.

Therefore, this paper aims to unify VOS and VOT and improve above shortcomings by integrating boxes and masks in a multi-object framework as Multi-object Integrated Tracking and Segmentation (MITS), as shown in Figure 1. For compatibility problem, a unified identification module is proposed to take both reference boxes in VOT and masks in VOS for initialization. The unified identification module encodes the reference boxes or masks into the unified identification embedding by assigning identities to objects. The coarse identification embedding from boxes is further refined to mitigate the gap between mask and box initialization. The unified identification module is more convenient than borrowing an extra model because it is trained with the whole model in an end-to-end manner.

Besides, the novel pinpoint box predictor is proposed for joint training and prediction with the mask decoder. Previous corner head or center head estimates a box by corners or a center point, which are not have to be inside the object. Emphasizing exterior points may distract learning target-oriented features and affect the mask prediction. To address this problem, we estimate the box by localizing pinpoints, which are always on the edge of the object. However, directly supervise the learning of pinpoints is infeasible due to the lack of annotation. Therefore we perform decoupled aggregation on the pinpoint maps and determine the box only by side-aligned pinpoint coordinates.

All the modules in our framework are not only compatible with two tasks, but also able to process multiple objects simultaneously. The multi-object training and prediction make our framework efficient and robust under complex scenes with multiple objects. Extensive experiments are conducted to demonstrate the strong compatibility and capacity of our framework. Experimental results show that our framework achieves SOTA performance on VOT benchmarks including LaSOT [23], TrackingNet [57] and GOT-10k [32], and VOS benchmark YouTube-VOS [84]. Our method improves 6% over previous SOTA VOT method on GOT-10k, and significantly improves the performance of box initialization on VOS benchmarks. In summary, our contributions are:

- We present a multi-object framework integrating boxes and masks for unified tracking and segmentation.
- The unified identification module is proposed to accept both masks and boxes for initialization.
- A novel pinpoint box predictor is proposed for accurate box prediction together with the mask decoder.

2. Related Work

Visual Object Tracking. In our context, we use visual object tracking (VOT) as a union of single object tracking (SOT) and multi-object VOT, preventing confusion with multi-object tracking (MOT) [56] which mainly considers object association between two detected object sets. VOT has been well studied in recent years. Correlation filter based methods [5, 30, 16, 14, 49, 3, 15, 53] train a correlation filter on training features and perform convolution on test features to predict classification scores for the target object. Siamese approaches [2, 70, 36, 35, 26, 75] use a Siamese network to learn a offline similarity metric between the template and search region, and perform cross-correlation to localize the target object. Recently some works [98, 19, 8, 77, 87, 69] adopt transformers [71] in visual tracking for feature extraction and correlation modeling [83, 13, 97, 7], which achieve promising performance. However, most VOT studies mainly focus on SOT, while ours is the first multi-object VOT framework.

Video Object Segmentation. In video segmentation tasks [39, 41, 38, 12], semi-supervised video object segmentation aims to track given objects with masks rather than boxes in VOT. Online VOS methods [6, 74, 51, 89, 55] fine-tune a segmentation model on the reference mask for each video, while matching-based VOS methods [68, 59, 80, 72, 92, 94] measure the pixel-level similarity to segment target objects. Space-time memory network (STM) [60] leverages a memory network to read object information from past frames with predicted masks, and is further improved and extended by many following works [46, 31, 66, 10, 93, 99]. Since masks are more consuming to annotate than boxes, VOS benchmarks [84, 63] usually have shorter videos than SOT benchmarks [23, 57]. Some methods [40, 42, 9] have been proposed for long-term VOS, but they lack consideration for VOT. AOT [93, 96] and its following works [95, 85, 12] realize simultaneous processing of multiple objects in VOS by the multi-object identification mechanism. We adopt the mechanism in our framework and further improve it for unified tracking and segmentation.

Multi-Object Tracking and Segmentation. MOT [56]/MOTS [73] is quite different from VOT/VOS, and the former relies on object detectors trained on *pre-defined* object categories, and focuses on the association of detected results. Although some work [81, 86] unified MOT/MOTS and VOT/VOS, we recognize the large gap between these two types of tasks and consider VOT and VOS that can generalize to *arbitrary* objects as the scope of our work.

Unified Tracking and Segmentation. Prior unified tracking and segmentation methods are listed in Table 1. Wang *et al.* developed a unifying approach SiamMask

Method	Initialization Box	Initialization Mask	Prediction Box	Prediction Mask	Extra Model	Multi-Object
SiamMask [78]	✓	✗	✓	✓	-	✗
D3S [48]	✓	✓	✗	✓	-	✗
SiamR-CNN [75]	✓	✗	✓	✗	Box2Seg [47]	✗
UniTrack [81]	✓	✓	✓	✓	-	✗
Unicorn [86]	✓	✓	✓	✓	-	✗
RTS [62]	✗	✓	✗	✓	STA [101]	✗
MITS (Ours)	✓	✓	✓	✓	-	✓

Table 1. We compare different prior unified approaches for unified visual object tracking and segmentation. For each method, we consider whether it offers native support for box and mask initialization and prediction without any extra model, and whether it is a multi-object framework.

[78] by augmenting previous tracking methods [2, 36] with a segmentation loss. D3S [48], a segmentation-centric tracker, applies two target models with geometric properties. The approximate mask is first estimated by a self tracking iteration for box initialization. Siam R-CNN [75] performs tracking by re-detection [25]. An extra off-the-shelf Box2Seg [47] network is needed to generate masks from boxes for the segmentation task. Alpha-Refine [88] is a plug-and-play model that refines coarse predictions from a base tracker. It provides good practice for refining tracking results by mining detailed spatial information.

UniTrack [81] uses a shared backbone for five different tracking related tasks (SOT, VOS, MOT, MOTS, PoseTrack [1]), while its main topic is the self-supervised pre-trained representation for these tasks. Unicorn [86] unifies four tracking tasks (SOT, VOS, MOT, MOTS) by using the target prior in the unified input head, and sets up two branches for box and mask output. Differently, we only considered VOT and VOS with strong generalization ability for compact unification. RTS [62] performs robust tracking by augmenting a VOS method [4] with an instance localization branch. An extra STA [101] network is borrowed to convert a box annotation to a mask for initialization in VOT. Compared with it, we follow the mask-oriented modeling but our framework supports multi-object, more compatible in input and output, and also achieves higher performance.

3. Method

3.1. Preliminary

Multi-Object Identification Mechanism. The multi-object identification mechanism is proposed in the VOS method AOT [93] and used in its following work [91, 95]. Information about all the target objects in the reference masks is encoded into an IDentification Embedding by an IDentity Bank. In detail, the ID bank stores M learnable high dimensional ID vectors, and each ID is responsible for representing one object. The ID embedding is obtained by assigning ID vectors to N objects ($N \leq M$) according to the reference mask. We adopt this mechanism in our framework, and further improve it by integrating both box and mask representation in a Unified IDentification Module, which achieves unified initialization for VOS and VOT.

Pinpoint Definition & Properties. *Pinpoints* are defined as the contiguous points between an object and its axis-aligned rectangle bounding box, which pin the box tightly around the object. Pinpoints are first proposed as *Extreme Points* for efficient object annotation instead of boxes [61]. We use the name *Pinpoints* in our paper to emphasize the connection constructed by pinpoints between the box and mask of an object, and further explore its properties for box head design.

Property 3.1 (Pinpoint Existence) *Given an object and its axis-aligned rectangle bounding box, there is at least one pinpoint on each side of the box. (An example for multiple pinpoints on one box side is in Figure 5.)*

Property 3.2 (Pinpoint-Based Box) *An axis-aligned rectangle bounding box can be fully determined by four parameters, which are the side-aligned coordinates of four pinpoints on four sides respectively, i.e. the y coordinates of top and bottom side pinpoints and the x coordinates of left and right side pinpoints (Illustrated in Figure 4 and 5).*

3.2. Pipeline Overview

Our unified pipeline for VOT and VOS tasks can be formulated as encoding-propagation-decoding (Figure 2). There are two beliefs for constructing our unified framework. The first is *mask-oriented*, i.e. extracting and maintaining detailed spatial information as much as possible, which has been proved to be more robust in prior work [48, 88, 62]. The second is *multi-object mask-box integrated*, i.e. natively supporting simultaneously processing multiple boxes or masks from input to output, which has never been achieved by prior work.

Unified Box-Mask Encoding. At the beginning, a reference frame or memory frame $I_m \in R^{H_I \times W_I}$ is encoded by a backbone encoder (which shares weights for all frames) into the visual embedding $X_m \in R^{HW \times C}$. The reference masks $Y_m \in R^{HW \times N}$ or boxes for N objects are encoded by the Unified Identification Module (Section 3.3) into the Unified IDentification Embedding $E_{id} \in R^{HW \times C}$. Then, the visual embedding is sent to the propagation module to obtain the memory embedding for propagation, together with the ID embedding. Note that any frame with predicted boxes or masks can become a memory frame, and the memory embedding is extended from $R^{HW \times C}$ to $R^{THW \times C}$ with T memory frames.

Spatial-Temporal Propagation. For the current frame, the visual embedding obtained from the encoder is fed into the propagation module. In the propagation module, the embedding of current frame is transformed into the query

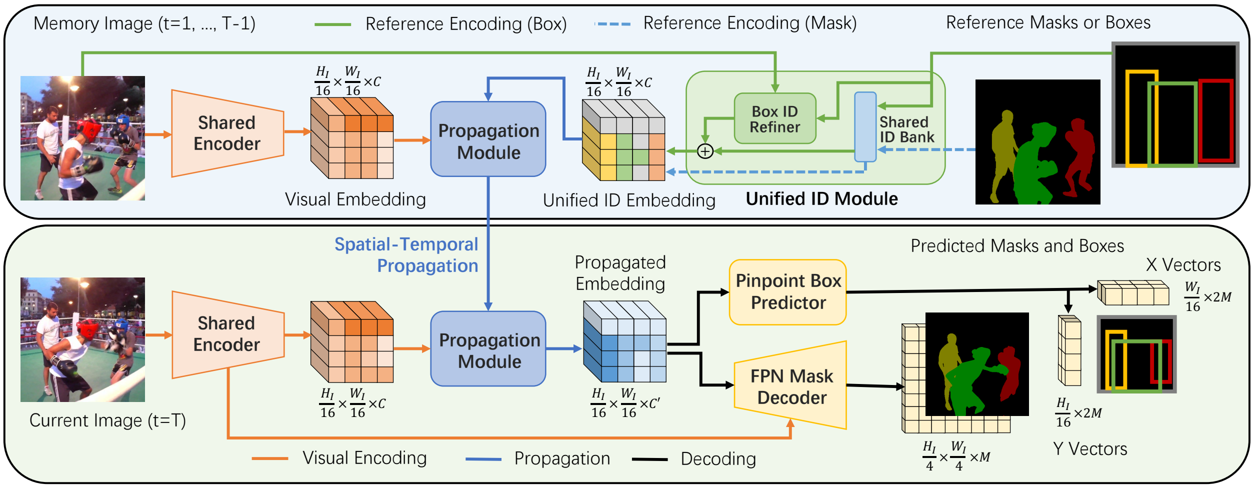

Figure 2. The unified encoding-propagation-decoding pipeline in MITS for VOT and VOS task. Both box and mask reference can be encoded by the unified identification module. The unified ID embedding is propagated by propagation module together with the visual embedding from past frames to the current frame. The propagated embedding is decoded into boxes and masks by separate branches.

$Q_t \in R^{HW \times C}$ and the memory embedding is transformed into the key $K_m \in R^{THW \times C}$ and value $V_m \in R^{THW \times C}$. The ID embedding $E_{id} \in R^{THW \times C}$ is fused with V_m and propagated from memory frames to the current frame via attention mechanism [71, 60, 93, 95]:

$$Attn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}}) \cdot V, \quad (1)$$

$$V_t = Attn(Q_t, K_m, V_m + E_{id}). \quad (2)$$

Dual-Branch Decoding. Finally, the embedding after propagation is passed to two decoding branches for different prediction. The decoder for predicting masks is a feature pyramid network (FPN) [44]. While the decoder for predicting boxes is a transformer-based pinpoint box head (Section 3.4). Both of these two decoders predict masks or boxes for all objects together. For the mask branch, the predicted logits $S_m \in R^{H_s W_s \times M}$ are for object capacity M, and the probability $P_m \in R^{HW \times N}$ for actual N objects is obtained by using $Softmax$ on N channels of M. For the box branch, the predicted probability vectors $P_b^{x_1,x_2} \in R^{W \times 2M}$ and $P_b^{y_1,y_2} \in R^{H \times 2M}$ form M boxes. We choose N boxes from M predicted boxes.

3.3. Unified Identification Module

For mask-based VOS models, directly taking a box for initialization will lead to severe performance degradation [75, 62], as shown in Table 2. A choice is to borrow an extra model which generates masks from boxes [62]. We argue such a strategy has two shortcomings. First, it is cumbersome train another extra model. Second, the extra model can't be optimized together with the VOS model. To address these problems, we propose a Unified IDentification Module (UIDM), which has three advantages: 1) It supports both box and mask initialization, keeping rich details in the masks while refining coarse information in the boxes. 2) It encodes the reference in a multi-object way, all objects are integrated in one embedding. 3) It can be directly trained as a part of the whole model in an end-to-end manner.

Box-Mask Integrated Identification. The UIDM module consists of two parts, the shared ID bank and the Box ID Refiner (BIDR), as shown in Figure 3. If the reference is mask representation, the shared ID bank generates an ID embedding directly from the masks, which stores rich object information. If the reference is box representation, the boxes are first converted to box-shaped masks, and the shared ID bank generates a coarse ID embedding. Then the coarse ID embedding is refined by BIDR into a fine ID embedding with more object details, which is unified with the mask reference:

$$E_{id} = \begin{cases} ID(Y_m), & \text{if } Y_m \text{ is mask ref.} \\ ID(Y_m) + R(I_m, Y_m), & \text{if } Y_m \text{ is box ref.} \end{cases} \quad (3)$$

Box ID Refinement. To extract accurate multi-object information from the image according to the boxes, BIDR is designed as a dual path transformer with both self-attention and cross attention layers. One is the image path for global information collection, the other is the object path for local information extraction. First, self-attention is used in both paths. The image path can learn the global information such positions and shapes around whole image including all objects, while the object path can learn detailed information for each object and interactive information between objects such as occlusion. Then, the dual cross attention is used for

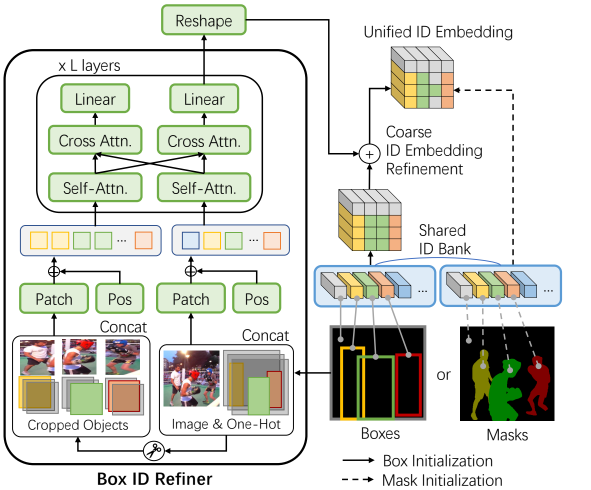

Figure 3. Illustration of the unified identification module, which consists of the shared ID bank and the box ID refiner. The procedures of box initialization and mask initialization are both shown.

the information exchange between two paths:

$$V_O' = Attn(Q_I, K_O, V_O),\ V_I' = Attn(Q_O, K_I, V_I) \quad (4)$$

where the subscript I is for the image path and O is for the object path. The local information is integrated into the global image path, which helps to generate finer ID embedding. Meanwhile, the global image information also flows to local objects in the boxes, helping to get better object discriminative information, which is favorable in multi-object scenario. After several transformer layers, the refinement embedding can represent more accurate information about the target objects than boxes. The output of BIDR is added with the coarse ID embedding to get a unified fine ID embedding, and then sent to propagation.

Auxiliary Mask Reconstruction. Although the shared ID bank and BIDR in UIDM can be trained in an end-to-end manner, direct learning how to refine a high dimensional embedding may be awkward. We mitigate the problem by employing an extra ID decoder during training. The UIDM acts as an encoder to generate the latent ID embedding, and the ID decoder reconstructs the mask from the latent ID embedding. In this way, the training of UIDM gets easier and more effective. Note that the ID decoder is not used in test.

3.4. Pinpoint Box Predictor

Keypoint-based box head is popular in detection and tracking, such as the corner head [34, 87, 13] and the center head [20, 97]. The corner head predicts top left and bottom right points while the center head predicts a box center and box sizes in x and y directions. The corners are exterior points to the objects (even the center does not have to be inside the object). Emphasizing exterior points distracts

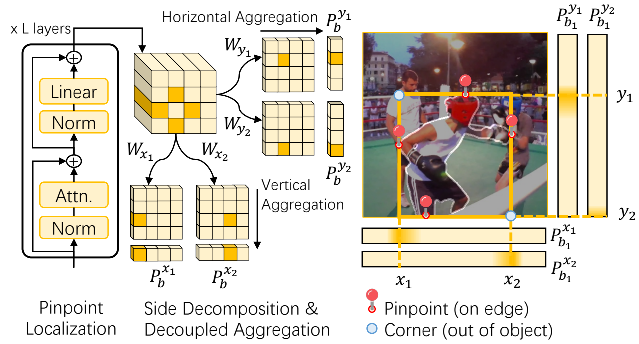

Figure 4. Illustration of pinpoint-based box prediction. The pinpoints are first localized by a transformer, and then projected into four probability maps for four sides. Then the probability maps are aggregated horizontally and vertically to extract side-aligned coordinates of pinpoints on each side.

the learning of accurate object-centric representation, especially when it is shared by both box and mask branch.

To make the box predictor harmonious in representation learning with the mask decoder, we propose a pinpoint box head, which determine a box of an object by localizing its pinpoints, illustrated in Figure 4 and visualized in Figure 5. A prior method [102] has attempted to use pinpoints for object detection, but it still relies on center point for object localization, and requires pinpoint annotations for supervision. However, it is hard to annotate all possible pinpoints for supervision, for example a side-aligned object edge includes infinite pinpoints. To solve this, we propose a decoupled aggregation strategy, which eases the demand for pinpoint annotation for direct supervision.

Pinpoint Localization. Given the propagated embedding $D_t \in R^{HW \times C'}$ as the output from the propagation module, it is fed into several transformer layers with self-attention [71] to perform pinpoint localization. With the powerful global modeling capacity of attention mechanism, object edge features are recognized. Later, pinpoints can be localized by exploring the extreme positions of the edges.

Decoupled Aggregation. A box has four or more pinpoints according to Property 3.1, while it can be determined by only four parameters, so it's redundant to directly predict all the pinpoints to obtain the box. In addition, there is no pinpoint annotation for direct supervision in our training data. Based on Property 3.2, we propose a decoupled aggregation strategy to predict the box by extracting the side-aligned coordinates from pinpoints on four sides. First, we decompose the pinpoints on the top, bottom, left and right sides from the localization feature maps by convolution layers $W_{x_1,y_1,x_2,y_2} \in R^{C' \times 4M}$ into four score maps. The score maps are normalized by $Softmax$ into probability maps. Then we horizontally aggregate top/bottom pinpoint probability maps, and vertically aggregate left/right

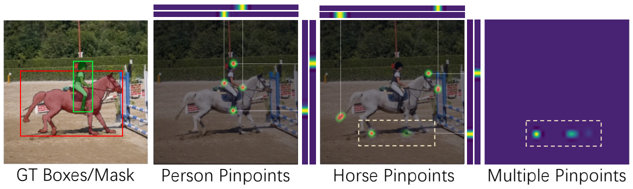

Figure 5. Visualization of pinpoint localization and decoupled aggregation. The scene includes two targets, a person and a horse. There are multiple approximate pinpoints on the bottom side of the horse, with each one localized on one hoof.

pinpoint probability maps into probability vectors $P_b^{x_1,x_2} \in R^{W \times 2M}$, $P_b^{y_1,y_2} \in R^{H \times 2M}$:

$$P_b^{x_1,x_2} = \sum_y Softmax(W_{x_1,x_2} \cdot SA(D_t)), \quad (5)$$

$$P_b^{y_1,y_2} = \sum_x Softmax(W_{y_1,y_2} \cdot SA(D_t)). \quad (6)$$

where SA is self-attention. Then the boxes for all N objects are determined by $[\mathbf{x_1}, \mathbf{x_2}, \mathbf{y_1}, \mathbf{y_1}] \in R^{M \times 4}$, which are predicted by $soft-argmax$:

$$\mathbf{x_1^T} = C_x \cdot P_b^{x_1},\ \mathbf{x_2^T} = C_x \cdot P_b^{x_2} \quad (7)$$

$$\mathbf{y_1^T} = C_y \cdot P_b^{y_1},\ \mathbf{y_2^T} = C_y \cdot P_b^{y_2} \quad (8)$$

where $C_x = [0, 1, ..., W]$, $C_y = [0, 1, ..., H]$ are the coordinate arrangements in x and y directions. Note that we select corresponding N boxes for N objects from M.

Box-Mask Collaboration. Although a single mask decoder can predict masks and convert them to bounding boxes, there are two advantages of employing both a mask decoder and a box predictor: *box-mask synergy* in training and *box-mask consistency* during test. We find training two branches together yields better performance than training a single mask branch (Table 5), which proves the pinpoint box branch can promote the learning of better edge-aware representation of objects. In addition, the box representation achieves higher precision than the mask representation on VOT benchmarks. During test, we can also calculate the consistency of the two branches by the box IoU, which can be used to indicate the reliability of the results.

3.5. Loss and Optimization

The total loss consists of mask loss and box loss. For mask branch, the loss functions we use are Cross Entropy loss and Jaccard loss, and for box branch we use L1 loss and Generalized IoU loss [64]. For the ID decoder, we only use Cross Entropy loss to supervise the mask reconstruction.

In our framework, both the mask Y_0^{mask} and box Y_0^{box} reference can be used for initialization. The target is to find parameters θ of the model f for two objectives:

$$\min_\theta L(f_\theta(X_0, ..., X_{t-1}, X_t, Y_0^{mask}), Y_t), \quad (9)$$

$$\min_\theta L(f_\theta(X_0, ..., X_{t-1}, X_t, Y_0^{box}), Y_t). \quad (10)$$

The output of the model $\hat{Y}_t$ and the ground truth Y_t include both masks and boxes, but here we simplify them as a single term. The model needs to achieve the minimal total loss both for box and mask initialization. To solve this, we randomly pick one objective and optimize it in each step and alternate between two objectives. The probability of choosing the box format for initialization is set as 0.3.

4. Experiment

4.1. Implementation Details

Model Setting. The backbone encoder of MITS is ResNet-50 [29]. For transformers in UIDM, propagation module and pinpoint box head, we all set 3 layers. The propagation module can be flexible in our framework, and we use the carefully designed gated propagation module proposed in DeAOT [95] by default. For VOT, we use the box branch for box evaluation by default, and the masks can also be evaluated by taking their bounding boxes.

Training. We use the same training data as RTS [62] for fair comparison, including the training sets of LaSOT [23], GOT-10k [32], Youtube-VOS [84] and DAVIS [63]. For VOT datasets, masks are estimated from boxes by STA [101] as in RTS [62]. The backbone encoder, propagation module and mask decoder are pre-trained on static images [22, 45, 11, 67, 27], following prior VOS methods [93, 95], while UIDM and the box predictor are trained from scratch. During training, we use 4 Nvidia RTX3090 GPUs, and the batch size is 16. The model is trained 100,000 steps with an initial learning rate of 2×10^{-4}. The learning rate gradually decays to 1×10^{-5} in a polynomial manner [92]. We use only one model to evaluate on all benchmarks, except GOT-10k test set, where only the training set of GOT-10k is used for its one-shot evaluation.

4.2. Evaluation on Single Object Tracking

LaSOT & TrackingNet. LaSOT [23] is a long-term tracking benchmark consisting of 280 videos in test set. The videos are at 30 FPS and have an average length around 2.5k frames. TrackingNet [57] is a short-term tracking benchmark, which contains 511 videos for test, and the average length is around 0.45k frames. The evaluation metrics are Success Rate (or AUC, Area Under Curve), Precision (P) and Normalized Precision (P_N).

We select very strong SOT methods for comparison in Table 2, such as MixFormer [13] and OSTrack [97], where

	LaSOT [23]			TrackingNet [57]			GOT-10k [32]		
SOT Method	AUC	P_N	P	AUC	P_N	P	AO	$SR_{0.5}$	$SR_{0.75}$
SiamFC [2]	33.6	42.0	33.9	57.1	66.3	53.3	34.8	35.3	9.8
MDNet [58]	39.7	46.0	37.3	60.6	70.5	56.5	29.9	30.3	9.9
SiamPRN++ [35]	49.6	56.9	49.1	73.3	80.0	69.4	51.7	61.6	32.5
DiMP [3]	56.9	65.0	56.7	74.0	80.1	68.7	61.1	71.7	49.2
MAMLTrack [76]	52.3	-	-	75.7	82.2	72.5	-	-	-
Ocean [100]	56.0	65.1	56.6	-	-	-	61.1	72.1	47.3
TrDiMP [77]	63.9	-	61.4	78.4	83.3	73.1	67.1	77.7	58.3
TransT [8]	64.9	73.8	69.0	81.4	86.7	80.3	67.1	76.8	60.9
STARK [87]	67.1	77.0	-	82.0	86.9	-	68.8	78.1	64.1
KeepTrack [54]	67.1	77.2	70.2	-	-	-	-	-	-
SBT [83]	66.7	-	71.1	82.2	87.5	-	70.4	80.8	64.7
ToMP [53]	68.5	79.2	73.5	81.5	86.4	78.9	-	-	-
MixFormer [13]	70.1	79.9	76.3	83.9	88.9	83.1	70.7	80.0	67.8
AiATrack [24]	69.0	79.4	73.8	82.7	87.8	80.4	69.6	80.0	63.2
OSTrack [97]	71.1	81.1	77.6	83.9	88.5	83.2	73.7	83.2	70.8
SwinTrack [43]	69.6	78.6	74.1	82.5	87.0	80.4	69.4	78.0	64.3
Unified Method									
SiamMask [78]	-	-	-	72.5	77.8	66.4	51.4	58.7	36.6
D3S [48]	-	-	-	72.8	76.8	66.4	59.7	67.6	46.2
SiamR-CNN [75]	64.8	72.2	-	81.2	85.4	80.0	64.9	72.8	59.7
Unicorn [86]	68.5	76.6	74.1	83.0	86.4	82.2	-	-	-
RTS [62]	69.7	76.2	73.7	81.6	86.0	79.4	-	-	-
MITS (Ours)	72.0	80.1	78.5	83.4	88.9	84.6	80.4	89.8	75.8

Table 2. Evaluation results on SOT benchmarks. There are two groups of methods 1) SOT only methods, 2) unified tracking and segmentation methods. Specially, one-shot setting is followed on GOT-10k test. The best two results are shown in red and blue.

large-scale transformers [18, 82] pretrained on ImageNet-22k [17] or by MAE [28] are used. Despite MITS only uses a ResNet-50 [29] backbone, it still achieves highest Success Rate/Precision on LaSOT and highest Normalized Precision/Precision on TrackingNet. Although MITS searches objects in the whole image, it runs at 20 FPS, which is comparable with MixFormer. MITS also significantly improve the performance of unified methods such as Unicorn [50], by 3.5% Success Rate and 4.4% Precision on LaSOT. MITS achieves impressive Precision owing to the accurate localization and robust decoupled prediction of the pinpoint head. Qualitative results are shown in Figure 6. We select video sequences which have multiple similar objects, and MITS tracks the target more robustly than other SOT methods, benefited from the multi-object framework.

GOT-10k. GOT-10k [32] is a large-scale short-term tracking benchmark, offering 9.34k videos for training, 180 videos for validation and 420 videos for test. We follow the *one-shot protocol* to verify the generalization ability of MITS on tracking, which requires GOT-10k training only to avoid biased evaluation results. MITS achieves absolutely superior performance on GOT-10k, with around 6% improvement over prior SOTA SOT method OSTrack [97]. The improvement can be attributed to two aspects. First, UIDM extracts detailed spatial information from boxes, which leads to accurate initialization. Second, MITS propagates the detailed spatial information from multiple memory frames with a long temporal coverage, and the prediction is on the whole image. Considering the videos are at 10FPS, the impressive results suggest that MITS is capable of robustly tracking objects in low frame rate videos, where motion is faster than that in high frame rate videos. SOT methods usually have a strong temporal assumption (searching in a local area) and may fail to capture fast moving objects.

		YouTube-VOS 19 Val [84]					DAVIS 17 Val [63]		
VOS Method	Init.	$\mathcal{G}$	$\mathcal{J}/\mathcal{F}_{seen}$		$\mathcal{J}/\mathcal{F}_{unseen}$		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
STM [60]	M	79.2	79.6	73.0	86.3	80.6	81.8	79.2	84.3
CFBI [92]	M	81.0	80.6	85.1	75.2	83.0	81.9	79.1	84.6
SST [21]	M	81.8	80.9	-	76.7	-	-	-	-
HMMN [66]	M	82.5	81.7	86.1	77.3	85.0	84.7	81.9	87.5
CFBI+ [94]	M	82.6	81.7	86.2	77.1	85.2	-	-	-
JOINT [52]	M	82.8	80.8	84.8	79.0	86.6	83.5	80.8	86.2
STCN [10]	M	82.7	81.1	85.4	78.2	85.9	85.4	82.2	88.6
AOT [93]	M	85.3	83.9	88.8	79.9	88.5	84.9	82.3	87.5
XMem [9]	M	85.5	84.3	88.6	80.3	88.6	86.2	82.9	89.5
DeAOT [95]	M	85.9	84.6	89.4	80.8	88.9	85.2	82.2	88.2
Unified Method									
D3S [48]	M	-	-	-	-	-	60.8	57.8	63.8
SiamR-CNN [75]	M	-	-	-	-	-	74.8	69.3	80.2
Unicorn [86]	M	-	-	-	-	-	69.2	65.2	73.2
RTS [62]	M	79.7	77.9	75.4	82.0	83.3	80.2	77.9	82.6
MITS (Ours)	M	85.9	84.9	89.7	80.4	88.7	84.9	82.0	87.7
Box Init.									
SiamMask [78]	B	52.8	60.2	45.1	58.2	47.7	56.4	54.3	58.5
SiamR-CNN [75]	B	67.3	68.1	61.5	70.8	68.8	70.6	66.1	75.0
DeAOT [95]	B	68.2	67.9	69.6	64.0	71.3	64.0	61.6	66.4
RTS [62]	B	70.8	71.1	65.2	74.0	72.8	72.6	69.4	75.8
MITS (Ours)	B	81.8	81.9	86.2	75.2	83.7	81.1	77.5	84.7

Table 3. Evaluation results on VOS benchmarks. There are two groups of methods 1) VOS only methods, 2) unified tracking and segmentation methods. Different initialization formats, mask (M) or box (B), are compared separately. The best two results are shown in red and blue.

4.3. Evaluation on Video Object Segmentation

YouTube-VOS 2019 & DAVIS 2017. DAVIS [63] is a small-scale benchmark for VOS, which contains 30 videos with high quality annotation masks in 2017 validation set. YouTube-VOS [84] is a large-scale benchmark for VOS, including 507 videos in 2019 validation set with 65/26 seen/unseen object categories. The Region Jaccard $\mathcal{J}$ and the Boundary F measure $\mathcal{F}$ are used as metrics.

It can be found in Table 3 that the performance of most prior unified methods is not satisfactory compared with VOS only methods, and MITS achieves SOTA performance on YouTube-VOS. MITS also performs better than all previous unified methods on DAVIS 2017 validation set and also competitive with recent VOS methods [93]. Qualita-

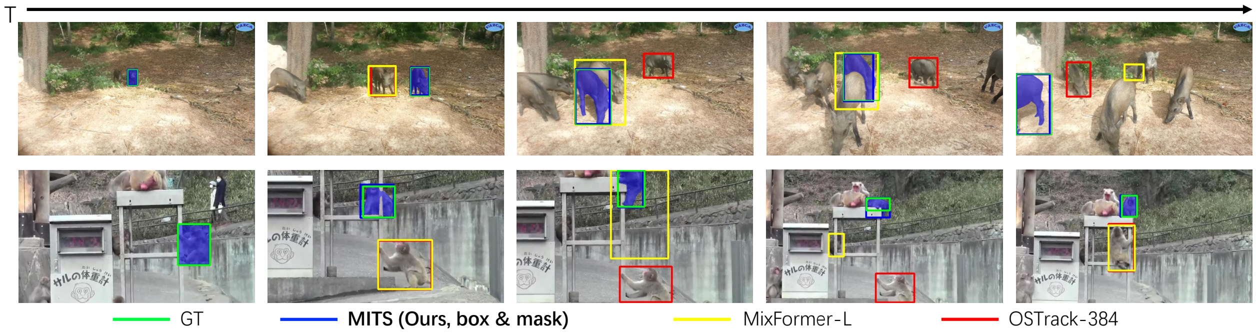

Figure 6. Qualitative results of MITS on VOT, compared with SOTA SOT methods MixFormer [13] and OSTrack [97].

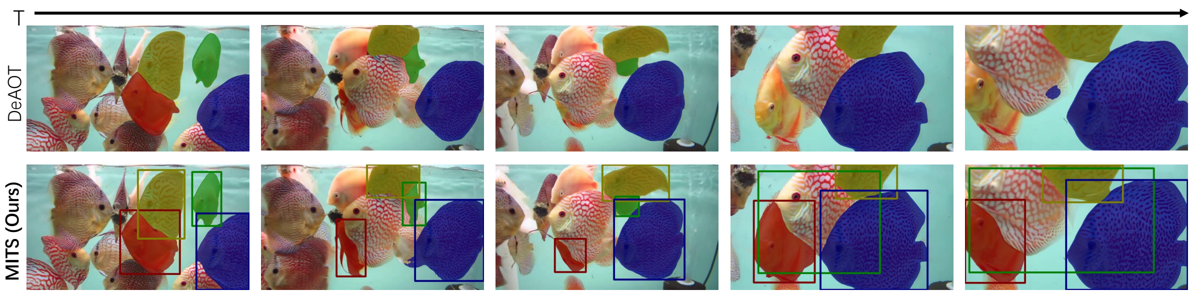

Figure 7. Qualitative results of MITS on VOS, compared with SOTA VOS method DeAOT [95]. For MITS, predictions from box predictor and mask decoder are both visualized.

tive results are shown in Figure 7.

Box Initialization. Following prior work [75, 62], we also tried using box initialization on VOS benchmarks. MITS with box initialization even outperforms prior unified methods with mask initialization on both YouTube-VOS and DAVIS. Besides, although DeAOT performs well on the VOS task, it relies heavily on fine masks for good results. If only boxes are available, it will suffer from sever performance degradation by around -20%. Differently, MITS only drops a small fraction from mask to box initialization, which shows the strong compatibility of UIDM.

5. Ablation Study

5.1. Unified Identification Module

Ablation results of UIDM are listed in Table 4. First, we compare MITS with DeAOT [95] on the tracking benchmark LaSOT [23]. Since DeAOT is a VOS method, we use off-the-shelf Alpha-Refine [88] model to generate masks from boxes for initialization. Compared with the separate models, our unified framework with UIDM achieves better performance. Second, we provide some ablation results of the box ID refiner (BIDR) in UIDM. During training, masks are reconstructed from the refined ID embedding (Mask Recon.), which leads to 0.5% Success Rate/1.7% Precision improvement. The dual cross attention (Dual CA) is used in BIDR to effectively exchange information between the

Method	Box-Mask Module	LaSOT [23] AUC	P_N	P
DeAOT [95]	None	65.9	70.7	68.4
DeAOT [95]	AlphaRefine [88]	67.5	72.6	71.1
MITS	UIDM	**72.0**	**80.1**	**78.5**
MITS	UIDM, no Mask Recon.	71.5	79.0	76.8
MITS	UIDM, no Dual CA	70.9	78.2	76.4

Table 4. Ablation of the unified identification module. We also compare MITS with SOTA VOS method DeAOT [95] with extra model AlphaRefine [88] on the VOT task.

global image and local cropped boxes. If only a single cross attention for the image path in each layer of BIDR is used, the Success Rate drops 1.1% and Precision drops 2%. The bidirectional information exchange is crucial in extracting detailed spatial information from boxes.

5.2. Pinpoint Box Predictor

Ablation results of the box predictor is in Table 5. We tried corner head, and other variants of the pinpoint head. Compared with a single mask branch, the mask branch is promoted by the box branch during training. Except explicitly predicting the probability maps of pinpoints, another choice is to localize the pinpoints implicitly in high dimension feature maps. The decoupled aggregation on probability maps is replaced by decoupled pooling on localized feature maps. The feature maps are pooled to vectors and then projected to probability vectors. The implicit pinpoint localization performs worse than explicit version on Suc-

Localization	Predict.	Box Branch AUC	Box Branch P_N	Box Branch P	Mask Branch AUC	Mask Branch P_N	Mask Branch P
None	None	-	-	-	71.5	78.8	77.2
CornerNet [34]	Corner	69.1	79.5	75.9	71.7	78.5	77.6
FPN [44]	Pinpoint*	70.6	78.9	77.3	71.2	78.8	78.4
Transformer [71]	Pinpoint*	71.1	79.9	78.4	72.0	79.9	**78.4**
Transformer [71]	Pinpoint	**72.0**	**80.1**	**78.5**	**72.3**	**79.9**	78.2

Table 5. Ablation results of the box predictor on LaSOT [23]. We tried corner head, and other pinpoint head variants like implicit pinpoint localization (marked by $*$).

Backbone	PM	Layer	LaSOT [23] AUC	LaSOT [23] P_N	LaSOT [23] P	FPS
ResNet-50 [29]	GPM [95]	3	**72.0**	**80.1**	**78.5**	20
ResNet-50 [29]	LSTT [93]	3	69.3	76.3	74.5	13
MobileNetV2 [65]	GPM [95]	2	67.8	74.6	72.6	**27**

Table 6. Results of different propagation modules (PM) and backbones in our framework.

cess Rate. In addition, we also tried aligning box predictor and mask decoder with the similar FPN [44], but the result is not as good as transformer [71].

5.3. Propagation Module

The propagation module can be flexible in our framework, and a general form has been shown in Equation 2. In practice, we use the carefully designed Gated Propagation Module (GPM) proposed in DeAOT [95] by default, and the number of GPM layer is 3. We also tried other variants including 2-layer GPM or 3-layer Long Short-Term Transformer (LSTT) proposed in AOT [93]. Readers can refer to AOT [93] and DeAOT [95] for more detailed structures of the propagation modules. As the experimental results show in Table 6, GPM performs better than LSTT on both efficiency and performance. Fewer layers of propagation module and a more lightweight backbone lead to the trade-off between efficiency and performance.

5.4. Training Data

There are two options for training data: (A) VOS box&mask + VOT box&pseudo-mask, (B) VOS box&mask + VOT box. (A) is taken as default and we directly borrow the generated pseudo-masks for VOT data from prior work [62]. While due to the strong compatibility of MITS, (B) is also originally supported for training. Table 7 shows that MITS trained with (B) can still achieve strong results. We use different dataset repeats to balance different sources.

5.5. Box and Mask Prediction

In our framework, we employ two branches, a mask decoder and a box predictor for joint training and prediction. Training with two branches yields better results than one branch, while they also have different prediction strategies. Previous work [62] has shown mask representation is more robust than box representation. However, we find the box representation has higher Precision while the mask representation has higher Success Rate, as shown in Table 8. A possible explanation is the different strategies of the two branches for multiple similar objects. If the model fails to tell the objects apart, box predictor often chooses one object but mask decoder often covers multiple possible objects.

	VOT training data Mask	VOT training data Box	TrackingNet [57] AUC	TrackingNet [57] P_N	TrackingNet [57] P	LaSOT [23] AUC	YTB19 [84] $\mathcal{G}$
(A)	✓	✓	83.4	88.9	84.6	72.0	85.9
(B)	✗	✓	82.8	87.9	83.1	70.7	85.7

Table 7. Comparison between different training data strategies. Pseudo-masks for VOT data is used in (A) but not in (B), with the mask loss disabled for VOT data.

Branch	LaSOT [23] AUC	LaSOT [23] P_N	LaSOT [23] P	TrackingNet [57] AUC	TrackingNet [57] P_N	TrackingNet [57] P
Box Predictor	72.0	**80.1**	**78.5**	83.4	**88.9**	**84.6**
Mask Decoder	**72.3**	79.9	78.2	**83.5**	88.5	84.0

Table 8. Comparison between box prediction and mask prediction on SOT benchmarks.

Another difference between two branches is on the prediction when the target object is absent. SOT benchmarks like LaSOT [23] ignore the frames when the object is absent (fully occluded or out of view), while VOS benchmarks like YouTube-VOS [84] requires to predict an empty mask with all background. Some SOT methods [13, 97] predicts boxes for all frames without considering whether the object is absent, while VOS methods need to precisely predict a background mask if the object is absent. As a unified framework, MITS can infer whether an object is absent with the help of the mask prediction, as shown in Figure 7.

6. Conclusion

In this paper, we propose a multi-object box-mask integrated framework for unified visual object tracking and segmentation. The proposed unified identification module shows strong compatibility on both box and mask initialization. Refining the ID embedding from boxes helps the model to capture more detailed spatial information about target objects. Besides, the pinpoint box predictor is proposed for accurate box prediction, considering the joint representation learning of both box and mask branches. The decoupled aggregation strategy realizes indirect supervision without any pinpoint annotation. The pinpoint head not only facilitates the target-oriented representation learning to benefit the mask branch, but also achieves higher precision than the mask branch. As a result, our framework achieves SOTA performance on both VOT and VOS benchmarks.

Acknowledgements. This work was supported by the Fundamental Research Funds for the Central Universities (No. 226-2023-00048, No. 226-2023-00051).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 3

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016. 2, 3, 7

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019. 2, 7

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

[5] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010. 2

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 2

[7] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: a simplified architecture for visual object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 375–392. Springer, 2022. 2

[8] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 2, 7

[9] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 640–658. Springer, 2022. 2, 7

[10] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2, 7

[11] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 6

[12] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 2

[13] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 2, 5, 6, 7, 8, 9

[14] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017. 2

[15] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2020. 2

[16] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 472–488. Springer, 2016. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[19] Fei Du, Peng Liu, Wei Zhao, and Xianglong Tang. Correlation-guided attention for corner detection based visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6836–6845, 2020. 2

[20] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 5

[21] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5912–5921, 2021. 7

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[23] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling.

Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1, 2, 6, 7, 8, 9

[24] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 146–164. Springer, 2022. 7

[25] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[26] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6269–6277, 2020. 2

[27] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 6

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 7

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 9

[30] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 2

[31] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4144–4154, 2021. 2

[32] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 1, 2, 6, 7

[33] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, et al. The tenth visual object tracking vot2022 challenge results. In *European Conference on Computer Vision*, pages 431–460. Springer, 2022. 1

[34] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 5, 9

[35] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019. 2, 7

[36] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 2, 3

[37] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv preprint arXiv:2304.09854*, 2023. 1

[38] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube baseline for universal video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[39] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18857, 2022. 2

[40] Yu Li, Zhuoran Shen, and Ying Shan. Fast video object segmentation using the global context module. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 735–750. Springer, 2020. 2

[41] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10055–10069, 2023. 2

[42] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020. 2

[43] Liting Lin, Heng Fan, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *arXiv preprint arXiv:2112.00995*, 2021. 7

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4, 9

[45] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[46] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 661–679. Springer, 2020. 2

[47] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for

video object segmentation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV*, pages 565–580. Springer, 2019. 1, 3

[48] Alan Lukezic, Jiri Matas, and Matej Kristan. D3s-a discriminative single shot segmentation tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7133–7142, 2020. 2, 3, 7

[49] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6309–6318, 2017. 2

[50] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8781–8790, 2022. 7

[51] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018. 2

[52] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9670–9679, 2021. 7

[53] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8731–8740, 2022. 2, 7

[54] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13444–13454, 2021. 7

[55] Tim Meinhardt and Laura Leal-Taixé. Make one-shot video object segmentation efficient again. *Advances in Neural Information Processing Systems*, 33:10607–10619, 2020. 2

[56] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2

[57] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 1, 2, 6, 7, 9

[58] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302, 2016. 7

[59] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7376–7385, 2018. 2

[60] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2, 4, 7

[61] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, pages 4930–4939, 2017. 3

[62] Matthieu Paul, Martin Danelljan, Christoph Mayer, and Luc Van Gool. Robust visual tracking by segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 571–588. Springer, 2022. 2, 3, 4, 6, 7, 8, 9

[63] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2, 6, 7

[64] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[65] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 9

[66] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021. 2, 7

[67] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015. 6

[68] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2167–2176, 2017. 2

[69] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8791–8800, 2022. 2

[70] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2805–2813, 2017. 2

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser,

and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4, 5, 9

[72] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 2

[73] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019. 2

[74] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 2

[75] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020. 2, 3, 4, 7, 8

[76] Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6288–6297, 2020. 7

[77] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1571–1580, 2021. 2, 7

[78] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 2, 3, 7

[79] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David Crandall, and Luc Van Gool. A survey on deep learning technique for video segmentation. *arXiv e-prints*, pages arXiv–2107, 2021. 1

[80] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3978–3987, 2019. 2

[81] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021. 2, 3

[82] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 7

[83] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8760, 2022. 2, 7

[84] 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, 2, 6, 7, 9

[85] Yuanyou Xu, Zongxin Yang, and Yi Yang. Video object segmentation in panoptic wild scenes. In *International Joint Conference on Artificial Intelligence*, 2023. 2

[86] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 733–751. Springer, 2022. 2, 3, 7

[87] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021. 2, 5, 7

[88] Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5289–5298, 2021. 1, 3, 8

[89] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6499–6507, 2018. 2

[90] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 1

[91] Zongxin Yang, Xiaohan Wang, Jiaxu Miao, Yunchao Wei, Wenguan Wang, and Yi Yang. Scalable video object segmentation with identification mechanism. *arXiv preprint arXiv:2203.11442*, 2023. 3

[92] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020. 2, 6, 7

[93] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 4, 6, 7, 9

[94] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 7

[95] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances*

in Neural Information Processing Systems, 2022. 2, 3, 4, 6, 7, 8, 9

[96] Zongxin Yang, Jian Zhang, Wenhao Wang, Wenhua Han, Yue Yu, Yingying Li, Jian Wang, Yunchao Wei, Yifan Sun, and Yi Yang. Towards multi-object association from foreground-background integration. In *CVPR Workshops*, volume 2, 2021. 2

[97] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 341–357. Springer, 2022. 2, 5, 6, 7, 8, 9

[98] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6728–6737, 2020. 2

[99] Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, and Wenjun Zhang. Boosting video object segmentation via space-time correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2256, 2023. 2

[100] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 771–787. Springer, 2020. 7

[101] Bin Zhao, Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Generating masks from boxes by mining spatio-temporal consistencies in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13556–13566, 2021. 1, 3, 6

[102] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 850–859, 2019. 5