SOVTrack: Open-Vocabulary Multi-Object Tracking with Self-Supervised Pseudo Labeling and Feature Distillation

Zekun Qian[1,2], Ruize Han[3], Junhui Hou[2†], and Wei Feng[1]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
[3] Shenzhen University of Advanced Technology, Shenzhen, China
{clarkqian,wfeng}@tju.edu.cn, hanruize@suat-sz.edu.cn, jh.hou@cityu.edu.hk
https://github.com/zekunqian/SOVTrack

Abstract. Open-Vocabulary Multi-Object Tracking (OVMOT) represents a critical challenge requiring the detection and tracking of diverse object categories beyond predefined training classes, including both seen (base) and unseen (novel) categories in real-world videos. Existing approaches are fundamentally limited by their reliance on traditional closed-set tracking paradigms trained only on base classes and synthetic image pairs, failing to exploit the rich temporal dynamics inherent in video sequences for universal tracking capabilities. We present SOVTrack, a novel self-supervised framework that harnesses SAM2's universal representations to address general-purpose tracking from raw video data without manual supervision. Our approach introduces Dual-Direction Pseudo Labeling (DDPL) to automatically generate high-quality pseudo-labels through bidirectional temporal validation while dynamically assessing the difficulty of each training example, and employs a dual-branch architecture with Multi-Cue Adversarial Distillation (MCAD) to effectively transfer SAM2's multi-cue pixel-level universal tracking capabilities to object-level tracking scenarios while preserving its generalization power across diverse object categories. Experimental results demonstrate that SOVTrack achieves state-of-the-art performance using only 10K video frames (2% of existing methods' data requirements), delivering substantial improvements in both base and novel category tracking performance.

Keywords: Open-Vocabulary Tracking · Multi-Object Tracking · Self-Supervised Learning · Knowledge Distillation

1 Introduction

Multi-Object Tracking (MOT) has evolved from tracking simple categories like pedestrians and vehicles to addressing the diverse object landscape encountered

† Corresponding Author.

in real-world scenarios. This evolution has led to the development of Open-Vocabulary Multi-Object Tracking (OVMOT) [12], which extends traditional MOT to handle both seen (base) and unseen (novel) categories during inference. Unlike traditional MOT that focuses on tracking predefined categories encountered during training, OVMOT presents a more practical and challenging tracking paradigm that requires algorithms to demonstrate superior generalization capabilities across diverse object categories in the real world.

However, despite these demanding requirements for superior generalization, current OVMOT approaches face fundamental limitations in their training paradigms. Most existing methods [10, 12, 14] continue to follow traditional MOT training strategies, primarily focusing on base category objects and relying on appearance-based association features to associate objects across frames. This approach inherently limits their ability to generalize to novel categories, resulting in suboptimal tracking performance across the spectrum of object diversity encountered in open-vocabulary scenarios.

This generalization gap contrasts sharply with the success of open-vocabulary detection and segmentation [7,16,25,26,29], which largely stems from the powerful universal representations learned by models like CLIP [22], demonstrating remarkable transfer capabilities across diverse downstream tasks. However, *the key challenge in constructing effective OVMOT lies in replicating this 'CLIP moment' within the object association stage* , shifting from CLIP-style novel recognition to SAM2-style novel association, where effectively transferring such universal representations to enable robust temporal association for novel category objects remains critical. Recent work such as MASA [13] has begun this exploration by leveraging SAM [9] to segment image datasets and employing data augmentation techniques to create geometric deformations of segmented regions for pseudo-label construction, thereby training more generalizable OVMOT models.

Nevertheless, MASA exhibits two fundamental limitations. From a data-quality perspective, its approach of generating image pairs through image augmentation overlooks critical temporal dynamics, occlusions, and motion patterns commonly encountered in real videos, which are essential for training robust trackers. Moreover, its noise filtering mechanism relies solely on segmentation thresholds without incorporating any temporal consistency checks, making it difficult to guarantee the quality of data used for tracking training. Additionally, MASA lacks a difficulty assessment for the generated samples, preventing it from dynamically guiding the subsequent training and inference processes. From a training perspective, MASA's supervision is limited to target-level contrastive learning, which fails to leverage the richer representations already provided by SAM. Consequently, even with 500K image pairs for training, its performance gains remain limited.

To address these limitations, our work directly tackles both issues. First, for data quality, we build pseudo-labels from raw video clips, enforce bidirectional temporal consistency to prune unreliable trajectories, and attach difficulty scores that guide later tracking processes. Second, to overcome MASA's insufficient utilization of training samples, we introduce a novel dual-branch association

architecture that efficiently harnesses the full potential of the SAM model. While retaining the target-level contrastive learning, we further employ a Multi-Cue Adversarial Distillation (MCAD) method that effectively achieves feature-level distillation, enabling stronger performance with only 2% of MASA's data.

Specifically, we adopt SAM2 [23], a prompt-guided Video Object Segmentation (VOS) model that extends the image-based SAM [9] to video scenarios. Unlike SAM which is limited to single-image segmentation, SAM2 learns rich representations from multiple tracking cues including association, localization, and classification information across video sequences, providing valuable foundations for learning effective association features in tracking scenarios. However, as a VOS model, SAM2 operates at the pixel level for instance segmentation and requires input prompts as tracking initialization points. Moreover, SAM2 relies on a simple FIFO memory queue without explicit trajectory management such as track lifecycle control. These characteristics make it unsuitable for autonomous target-level tracking applications required in OVMOT. To address these challenges and harness SAM2's powerful capabilities, we propose SOVTrack, a novel self-supervised OVMOT framework that treats SAM2 traces as hypotheses accepted by cycle closure before distillation, including two key innovations.

First, we introduce Dual-Direction Pseudo Labeling (DDPL), a detection-guided strategy that automatically generates high-quality pseudo-labels for self-supervised training. DDPL begins by utilizing the tracker's detector to identify objects of interest in the scene, which serve as initialization prompts for SAM2. Since objects of interest detection inherently produces numerous false positives, effective filtering becomes crucial. Our approach caches intermediate results from forward VOS processes and subsequently uses these results as new starting points for backward VOS. By evaluating consistency relationships in tracking trajectories, we assess the reliability of detection prompts, effectively filtering noisy detections while providing confidence scores that guide subsequent adaptive feature fusion. Second, we develop a dual-branch association architecture that strategically separates appearance-based contrastive learning from comprehensive knowledge distillation. The contrastive branch continues traditional target-level dense similarity learning to capture appearance features. The other adversarial branch implements Multi-Cue Adversarial Distillation (MCAD), a novel strategy that effectively distills associative features between targets from SAM2's representations, which contain association, localization, and classification information. This enables each branch to specialize in its respective learning objective while facilitating effective knowledge transfer. To maximize the benefits of both branches, we design an adaptive feature fusion mechanism guided by the confidence scores learned from DDPL.

Experimental results demonstrate the effectiveness of our approach, achieving state-of-the-art performance with remarkable efficiency. Using only 10K video frames for knowledge distillation, a significantly smaller dataset than existing methods, our method delivers substantial improvements that far exceed other methods, including 7.8% improvement in novel category association (AssocA) and 5.7% improvement in overall tracking accuracy (TETA) compared to MASA,

establishing a new paradigm for self-supervised OVMOT training. The main contributions of this work include:
• A novel cycle consistency mechanism that automatically generates high-quality pseudo-labels for self-supervised OVMOT training by filtering unreliable pseudo-labels through bidirectional tracking validation, while providing confidence measures to guide adaptive branch feature fusion.
• A dual-branch association architecture that effectively separates and integrates appearance-based learning with foundation model knowledge distillation. The architecture incorporates Multi-Cue Adversarial Distillation to fully exploit association, localization, and classification information, and employs an adaptive fusion mechanism that dynamically balances both branches based on tracking difficulty.
• Comprehensive experimental results demonstrating that our approach achieves state-of-the-art OVMOT performance with significant improvements across both base and novel categories, validating the effectiveness of our self-supervised foundation model distillation framework even with minimal video data requirements.

2 Related Work

Open-Vocabulary Multi-Object Tracking. Open-Vocabulary Multi-Object Tracking (OVMOT) emerges to address the limitations of traditional MOT methods, formally introduced by OVTrack [12]. OVMOT extends traditional MOT to handle both seen (base) and unseen (novel) categories during inference. Subsequent works explore various approaches: MASA [13] leverages large-scale unlabeled image datasets for self-supervised learning, SLAck [14] incorporates multi-cue information, OVTR [10] introduces transformer-based architectures and TRACT [15] incorporates plugin architectures to further enhance performance. More recent studies further extend OVMOT from complementary perspectives: VOVTrack [18] introduces self-supervised training through cycle consistency, COVTrack [19] proposes a new multi-cue fusion framework to improve tracking stability, DOVTrack [20] develops a diffusion-based sample augmentation model, and GOVTrack [21] further brings generative modeling into OVMOT. Despite these advances, current OVMOT methods face two fundamental limitations. First, they fail to fully exploit the tracking universality required for open-vocabulary scenarios, typically training association models using base category supervision while neglecting universal association capabilities necessary for novel object tracking. Second, these methods do not effectively leverage the temporal continuity inherent in video data, relying on synthetic image pairs or sparse frame sampling that misses rich temporal dynamics essential for robust tracking. Our proposed method effectively addresses these issues by leveraging video data through self-supervised learning.

Segment and Track Anything Models. Recent segment and track anything models are built upon SAM [9], which demonstrates remarkable zero-shot segmentation capabilities but is limited to single-image tasks. Early video extensions like Track Anything [27] and Segment and Track Anything [4] combine

SAM with traditional VOS methods, but struggle with domain gaps. SAM2 [23] advances to native video understanding through streaming memory and training on 50.9K videos. Building upon the SAM series works, many approaches employ adapter or LoRA layers [2, 3] to improve performance in specific domains by adapting the model's knowledge [17, 30]. While VOS-based approaches excel at pixel-level segmentation, MASA [13] pioneers transferring SAM's capabilities to general-purpose object tracking using 500K segmented images and geometric augmentations. However, MASA's reliance on image pairs and target-level pseudo-labels provides limited knowledge transfer, failing to fully exploit SAM's feature representations. Our SOVTrack addresses these limitations through comprehensive knowledge distillation that leverages both target-level and pixel-level features from SAM2's representations, achieving superior performance with only 10K video frames compared to MASA's 500K images.

3 Proposed Method

3.1 Overview

OVMOT faces two fundamental challenges: (1) traditional MOT methods trained only on base categories fail to generalize to novel objects, and (2) existing approaches inadequately exploit the rich temporal dynamics inherent in video sequences for universal tracking capabilities. The SOVTrack framework addresses these challenges through a self-supervised approach that effectively transfers SAM2's universal representations to object-level tracking. The key insight is to leverage SAM2's comprehensive understanding of association, localization, and classification cues while maintaining the discriminative power necessary for robust association across diverse object categories. As illustrated in Figure 1, our framework consists of two core components that work synergistically to achieve superior OVMOT performance: **Dual-Direction Pseudo Labeling** automatically generates high-quality training data from raw video sequences through a principled bidirectional validation mechanism, while providing confidence measures to guide adaptive branch feature fusion. **Dual-Branch Association** separates two competing objectives through specialized branches sharing a ResNet-50 backbone. The *contrastive branch* uses contrastive loss for discriminative similarity learning with validated pseudo-labels. The *adversarial branch* distills SAM2 knowledge through three feature types: localization cues $\mathbf{f}_{\text{loc}}$ for spatial information, association cues $\mathbf{f}_{\text{assoc}}$ for temporal consistency, and classification cues $\mathbf{f}_{\text{cls}}$ for semantic understanding.

3.2 Dual-Direction Pseudo Labeling (DDPL)

The transition from SAM2's prompt-guided segmentation paradigm to autonomous OVMOT presents fundamental challenges in obtaining reliable object initialization without manual intervention. While SAM2 demonstrates exceptional temporal consistency when provided with accurate prompts, OVMOT requires

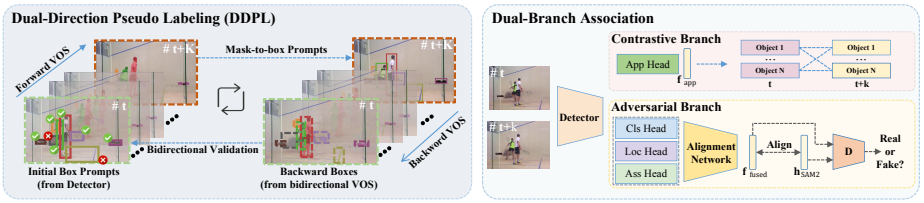

Fig. 1: SOVTrack framework overview. Left: DDPL employs bidirectional VOS validation for reliable pseudo-label generation. Right: Dual-Branch Association architecture with contrastive branch for appearance learning and adversarial branch for knowledge distillation, where fused features $\mathbf{f}_{\text{fused}}$ undergo adversarial training against SAM2 representations $\mathbf{h}_{\text{SAM2}}$ via discriminator D.

fully autonomous operation across diverse object categories. A critical first step in constructing high-quality pseudo-labels is to obtain reliable initialization prompts that can guide SAM2's VOS process effectively.

Detection-Guided Initialization Strategy. The most intuitive method for building pseudo-labels with SAM2 would be to directly segment the first frame and convert the resulting masks into bounding boxes as initialization prompts. While this approach appears simple and direct, it faces two fundamental problems that make it impractical for high-quality pseudo-label generation.

First, SAM2's segmentation results frequently exhibit cross-region errors, where a single mask incorrectly spans multiple distinct objects or object parts. This leads to bounding boxes of particularly low quality that fail to accurately localize individual objects, introducing significant noise into the training process. Second, in complex scenarios, especially under poor lighting conditions, SAM2's segmentation tends to be highly inaccurate and fragmented, producing numerous small, disconnected segments. This fragmentation dramatically increases false positive detections, making it extremely difficult to filter out low-quality pseudo-labels through subsequent validation processes. Representative examples of these issues are shown in red boxes in Figure 2.

To address these challenges, we propose a detection-guided initialization strategy that leverages the tracker's detector to identify objects of interest, which then serve as initialization prompts for SAM2. This approach offers two key advantages over direct SAM2 segmentation: (1) the detector provides higher-quality initial prompts by focusing on salient objects while filtering out background noise and fragmented regions, and (2) since the same detector is used during both pseudo-label generation and subsequent association training, this strategy ensures better adaptation and consistency throughout the learning pipeline.

However, even with detection-guided initialization, objects of interest detection inherently produces numerous false positives and inconsistent localizations, particularly when dealing with diverse object categories and complex scenes. This quality variation can severely compromise learning effectiveness if not properly addressed. As shown in the left part of Figure 1, to further ensure the quality of pseudo-labels generated from salient object detections, we introduce

Dual-Direction Pseudo Labeling (DDPL) validation, which employs bidirectional temporal consistency checks over detector proposals and SAM2 candidate traces to filter unreliable detections while providing confidence measures.

Forward-Backward VOS Validation. Given a video sequence $\mathcal{V} = \{I^t\}_{t=1}^{T}$ and the detector producing object proposals $\mathcal{D}^t = \{d_i^t\}_{i=1}^{N_t}$ for frame t, where each detection $d_i^t = (\mathbf{b}_i^t, s_i^t)$ consists of a bounding box $\mathbf{b}_i^t \in \mathbb{R}^4$ and confidence score s_i^t, we select high-confidence detections as initial prompts

$$\mathbf{P}^t = \{\mathbf{b}_i^t | d_i^t \in \mathcal{D}^t, s_i^t > \theta_{\text{det}}\}, \tag{1}$$

where θ_{det} is the detection confidence threshold.

Using these prompts, we perform forward VOS with SAM2 over K subsequent frames to obtain trajectories $\mathcal{T}^{\text{fwd}} = \{\tau_i^{\text{fwd}}\}_{i=1}^{|\mathbf{P}^t|}$, where each trajectory $\tau_i^{\text{fwd}} = \{m_i^{t+k,\text{fwd}}\}_{k=0}^{K}$ represents segmentation masks across consecutive frames starting from detection d_i^t.

To validate trajectory reliability, we extract bounding boxes from terminal masks and use them as initialization prompts for backward VOS

$$\mathbf{P}^{t+K} = \{\mathbf{b}(m_i^{t+K,\text{fwd}}) | \tau_i^{\text{fwd}} \in \mathcal{T}^{\text{fwd}}\}, \tag{2}$$

where $\mathbf{b}(\cdot)$ denotes bounding box extraction from segmentation masks.

The backward VOS produces trajectories $\mathcal{T}^{\text{bwd}} = \{\tau_i^{\text{bwd}}\}_{i=1}^{|\mathbf{P}^{t+K}|}$ from frame $t+K$ back to frame t. We then measure consistency between forward and backward trajectories through spatial overlap

$$\mathcal{S}_{\text{cycle}}^{i,t} = \text{IoU}(\mathbf{b}_i^t, \mathbf{b}(m_i^{t,\text{bwd}})), \tag{3}$$

where $\mathbf{b}_i^t$ is the original detection and $\mathbf{b}(m_i^{t,\text{bwd}})$ is the bounding box extracted from the backward VOS mask at frame t, where $m_i^{t,\text{bwd}}$ represents the segmentation mask of object i at frame t obtained through backward VOS. Box IoU matches box-based tracking and avoids accepting fragmented masks whose boxes drift. Trajectories with $\mathcal{S}_{\text{cycle}}^{i,t} > \theta_{\text{cycle}}$ are retained as reliable pseudo-labels, while inconsistent trajectories are filtered out.

Figure 2 visualizes the effectiveness of our forward-backward VOS validation. Many false positive detections exhibit background interference, incomplete coverage, or oversized bounding boxes. Our bidirectional validation effectively filters these unreliable detections by leveraging temporal consistency, ensuring only high-quality pseudo-labels contribute to training.

Pseudo-label Generation and Confidence Assessment. For validated trajectories, we systematically construct pseudo-labels that provide comprehensive supervision for both detection and association learning.

For each validated trajectory τ_i^{fwd} with masks $\{m_i^{t+k,\text{fwd}}\}_{k=0}^{K}$, we generate pseudo-labels for every object instance across all frames in the trajectory

$$\mathcal{T}_i^{\text{pseudo}} = \{(\mathbf{b}_i^{t+k}, \text{id}_i, \mathcal{C}_i^{t+k})\}_{k=0}^{K}, \tag{4}$$

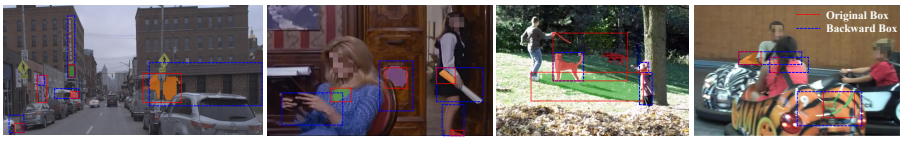

Fig. 2: Visualization of DDPL filtering results. Red boxes show problematic original detections, while blue dashed boxes represent backward boxes.

where $\mathbf{b}_i^{t+k} = \mathbf{b}(m_i^{t+k,\text{fwd}}) \in \mathbb{R}^4$ represents the bounding box extracted from the forward tracking mask at frame $t+k$, and id_i denotes the consistent object identity assigned across all frames in the trajectory.

Since we have obtained both forward tracking trajectories $\mathcal{T}^{\text{fwd}}$ and backward trajectories $\mathcal{T}^{\text{bwd}}$, we can compute spatial overlap at each time step to assess tracking difficulty. Note that unlike $\mathcal{S}_{\text{cycle}}^{i,t}$ (Eq. 3), which validates the initial detection prompt by checking the start-frame consistency, the per-frame confidence score $\mathcal{C}_i^{t+k}$ measures the agreement between forward and backward masks at each intermediate frame $t+k$:

$$\mathcal{C}_i^{t+k} = \text{IoU}(\mathbf{b}(m_i^{t+k,\text{fwd}}), \mathbf{b}(m_i^{t+k,\text{bwd}})), \tag{5}$$

where $m_i^{t+k,\text{bwd}}$ is the backward tracking mask at frame $t+k$. The confidence score $\mathcal{C}_i^{t+k}$ serves as a fine-grained indicator of per-frame tracking difficulty within each validated trajectory. High confidence scores indicate frames with stable appearance and motion, where traditional appearance-based features are sufficient. Lower scores indicate relatively more challenging frames where the adversarial branch's richer representations provide complementary benefits. This frame-wise confidence assessment provides complete supervision for the subsequent dual-branch learning processes.

3.3 Dual-Branch Association

Background on OVMOT Architecture. Existing OVMOT trackers [10,12,13] typically adopt a tracking-by-detection framework. Specifically, the detector first localizes objects and extracts feature embeddings $\mathbf{f}_r$ for each detected region r. These extracted features are then processed through three specialized heads: (1) a localization head that refines bounding box coordinates, (2) a classification head that predicts object categories, and (3) an appearance/association head that extracts discriminative features for temporal association. However, these methods predominantly rely on the single appearance/association head for object association, which limits their ability to effectively balance discriminative learning with generalization to novel categories.

Dual-Branch Architecture. To address this limitation, we propose a novel dual-branch balanced association architecture that strategically separates two competing objectives: learning discriminative appearance features for reliable object re-identification and acquiring comprehensive understanding for robust

generalization to novel categories. As illustrated in the right part of Figure 1, our architecture introduces two specialized branches, the *contrastive branch* and the *adversarial branch*, that share a ResNet-50 backbone of the detector but diverge in their learning strategies and feature representations.

Contrastive Branch. The contrastive branch contains a traditional appearance head that specializes in discriminative visual similarity learning using the validated pseudo-labels $\mathcal{T}_i^{\text{pseudo}}$ from DDPL. For each object instance i in the validated trajectories, the detector extracts feature embeddings $\mathbf{f}_r^{i,t+k}$ from the detected region. These features are then processed by the appearance head to obtain discriminative appearance features $\mathbf{f}_{\text{app}}^{i,t+k} \in \mathbb{R}^d$, and we construct similarity matrices between frames within the tracking window $\mathbf{S}_{\text{app}}^{t,t+k} = \mathbf{F}_{\text{app}}^{t}(\mathbf{F}_{\text{app}}^{t+k})^T$, where $\mathbf{F}_{\text{app}}^{t}$ and $\mathbf{F}_{\text{app}}^{t+k}$ represent the appearance feature matrices for frames t and $t+k$ respectively, with $k \in \{1, 2, \ldots, K\}$ denoting subsequent frames within the DDPL tracking window.

The contrastive branch is optimized using contrastive loss with the pseudo-label supervision

$$\mathcal{L}_{\text{app}} = -\sum_{i,j} y_{ij}^{\text{pseudo}} \log \frac{\exp(\mathbf{S}_{\text{app},ij}/\gamma)}{\sum_k \exp(\mathbf{S}_{\text{app},ik}/\gamma)}, \tag{6}$$

where y_{ij}^{pseudo} indicates whether objects i and j belong to the same trajectory according to $\mathcal{T}_i^{\text{pseudo}}$, and γ is the temperature parameter.

Adversarial Branch. Beyond generating high-quality pseudo-labels, DDPL also provides access to SAM2's rich universal representations that contain comprehensive multi-cue information. The adversarial branch leverages this opportunity to capture SAM2's extensive knowledge through sophisticated knowledge transfer. During the DDPL tracking process, we extract SAM2's internal object representations for every validated object instance. For each object i at frame $t+k$ within the tracking window, SAM2 generates object pointers through its cross-attention mechanism as $\mathbf{h}_{\text{SAM2}}^{i,t+k} = \text{CrossAttn}(\mathbf{f}_{\text{img}}^{t+k}, \mathbf{f}_{\text{prompt}}^{i}, \mathbf{f}_{\text{memory}}^{i,t+k})$, where $\mathbf{f}_{\text{img}}^{t+k}$, $\mathbf{f}_{\text{prompt}}^{i}$, and $\mathbf{f}_{\text{memory}}^{i,t+k}$ represent spatial-semantic features, localization guidance, and temporal consistency information respectively. The generated pointer $\mathbf{h}_{\text{SAM2}}^{i,t+k} \in \mathbb{R}^{d_{\text{SAM2}}}$ naturally integrates association, classification, and localization information.

To distill these rich representations, the adversarial branch extracts three types of cues from our tracker corresponding to SAM2's multi-source integration: localization cues $\mathbf{f}_{\text{loc}}^{i,t+k} \in \mathbb{R}^{d_{\text{loc}}}$ from the localization head for spatial information, association cues $\mathbf{f}_{\text{assoc}}^{i,t+k} \in \mathbb{R}^{d_{\text{assoc}}}$ from the implicit association head for temporal consistency, and classification cues $\mathbf{f}_{\text{cls}}^{i,t+k} \in \mathbb{R}^{d_{\text{cls}}}$ from the classification head for semantic understanding. Note that the implicit association head shares the same architecture as the appearance head but with different learned parameters, enabling specialized cue extraction for temporal association while maintaining architectural consistency.

These multi-cue features are then processed through our Alignment Network to generate representations compatible with SAM2's feature space. The Alignment Network first projects each feature type to a common dimensional space that matches SAM2's representation dimensionality $\mathbf{f}_*^{\text{proj},i,t+k} = \text{FC}_*(\mathbf{f}_*^{i,t+k}) \in \mathbb{R}^{d_{\text{SAM2}}}$, where $*$ denotes 'cls', 'loc', and 'assoc', respectively. Subsequently, the projected features are combined through a learnable fusion module that adaptively weighs the contribution of each feature type

$$\mathbf{f}_{\text{fused}}^{i,t+k} = \text{MLP}_{\text{fusion}}([\mathbf{f}_{\text{cls}}^{\text{proj},i,t+k}; \mathbf{f}_{\text{loc}}^{\text{proj},i,t+k}; \mathbf{f}_{\text{assoc}}^{\text{proj},i,t+k}]).$$

Multi-Cue Adversarial Distillation (MCAD). To enhance knowledge transfer quality, we introduce adversarial distillation where our Alignment Network acts as a generator creating SAM2-like representations, while a discriminator network evaluates their authenticity. The discriminator D is implemented as a three-layer MLP with LeakyReLU activations and dropout regularization. It takes feature vectors of dimension d_{SAM2} as input and outputs a scalar probability score indicating whether the input feature is real (from SAM2) or generated (from the Alignment Network).

The basic distillation loss encourages direct feature alignment between our fused features and SAM2's object pointers extracted from validated trajectories

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\mathcal{T}^{\text{pseudo}}|} \sum_{\mathcal{T}_i^{\text{pseudo}}} \sum_{k=0}^{K} \|\mathbf{f}_{\text{fused}}^{i,t+k} - \mathbf{h}_{\text{SAM2}}^{i,t+k}\|_2^2, \tag{7}$$

The discriminator D distinguishes between generated fusion features and genuine SAM2 features with the loss $\mathcal{L}_{\text{disc}} = -\mathbb{E}[\log D(\mathbf{h}_{\text{SAM2}}) + \log(1 - D(\mathbf{f}_{\text{fused}}))]$ where $\mathbf{h}_{\text{SAM2}}$ are real SAM2 features and $\mathbf{f}_{\text{fused}}$ are generated fusion features. The generator (Alignment Network) is trained adversarially to fool the discriminator with the loss $\mathcal{L}_{\text{adv}} = -\mathbb{E}[\log D(\mathbf{f}_{\text{fused}})]$.

The complete MCAD objective combines both components $\mathcal{L}_{\text{MCAD}} = \mathcal{L}_{\text{distill}} + \alpha\mathcal{L}_{\text{adv}}$, where α balances direct alignment with adversarial learning. The discriminator and alignment network are trained using an alternating optimization scheme to ensure stable adversarial training while maintaining effective knowledge distillation.

Adaptive Branch Feature Fusion. The effectiveness of our dual-branch architecture depends critically on intelligently combining features from both branches based on tracking conditions and object characteristics. While the contrastive branch excels at discriminative similarity learning for stable objects, the adversarial branch provides rich semantic and motion understanding crucial for challenging scenarios involving novel categories, occlusions, or complex motions.

Our adaptive fusion mechanism leverages the per-frame confidence scores $\mathcal{C}_i^{t+k}$ from DDPL to dynamically weight contributions from both branches. It is important to note that these scores are computed exclusively on validated trajectories that have already passed the cycle consistency filtering (Eq. 3), ensuring

that SAM2's tracking is globally reliable. Within these validated trajectories, $\mathcal{C}_i^{t+k}$ captures frame-level tracking difficulty: high scores suggest stable conditions where discriminative appearance features are sufficient, while lower scores indicate relatively more challenging frames (e.g., partial occlusions or deformations) where the richer semantic representations from the adversarial branch provide complementary cues beyond appearance alone.

From the fused features, we predict appearance reliability scores that measure current tracking difficulty

$$\hat{\mathcal{R}}_{\text{reliability}}^{i,t+k} = \text{sigmoid}(\text{FC}_{\text{pred}}([\mathbf{f}_{\text{cls}}^{\text{proj},i,t+k}; \mathbf{f}_{\text{loc}}^{\text{proj},i,t+k}; \mathbf{f}_{\text{assoc}}^{\text{proj},i,t+k}])), \tag{8}$$

which is supervised by the confidence scores from DDPL as

$$\mathcal{L}_{\text{reliability}} = \frac{1}{|\mathcal{T}^{\text{pseudo}}|} \sum_{\mathcal{T}_i^{\text{pseudo}}} \sum_{k=0}^{K} \|\hat{\mathcal{R}}_{\text{reliability}}^{i,t+k} - \mathcal{C}_i^{t+k}\|_2^2. \tag{9}$$

Since all training trajectories have been validated by DDPL, the predicted reliability score serves as a soft indicator of per-frame difficulty rather than an absolute measure of tracking failure. The final tracking features adaptively combine both branches accordingly

$$\mathbf{f}_{\text{final}}^{i,t+k} = \hat{\mathcal{R}}_{\text{reliability}}^{i,t+k} \mathbf{f}_{\text{app}}^{i,t+k} + (1 - \hat{\mathcal{R}}_{\text{reliability}}^{i,t+k}) \mathbf{f}_{\text{fused}}^{i,t+k}. \tag{10}$$

To ensure discriminative properties, we apply contrastive loss to the final fused features

$$\mathcal{L}_{\text{fusion}} = -\sum_{i,j} y_{ij}^{\text{pseudo}} \log \frac{\exp(\text{sim}(\mathbf{f}_{\text{final}}^{i,t+k_i}, \mathbf{f}_{\text{final}}^{j,t+k_j})/\gamma)}{\sum_l \exp(\text{sim}(\mathbf{f}_{\text{final}}^{i,t+k_i}, \mathbf{f}_{\text{final}}^{l,t+k_l})/\gamma)}. \tag{11}$$

3.4 Implementation Details

Overall Architecture. The complete training objective for the Alignment Network integrates all loss components: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{app}} + \mathcal{L}_{\text{MCAD}} + \mathcal{L}_{\text{reliability}} + \mathcal{L}_{\text{fusion}} + \mathcal{L}_{\text{loc}}$, where $\mathcal{L}_{\text{loc}}$ is the localization loss that leverages SAM2's high-quality segmentation masks by converting them to bounding boxes and supervising the tracker's localization head for improved localization accuracy. We employ the Smooth L1 loss from Faster R-CNN [24] for robust localization training. The discriminator is optimized separately using $\mathcal{L}_{\text{disc}}$ in the alternating training scheme. For fair comparison with existing methods, we adopt the same OV detector as OVTrack [12] and employ ResNet-50 as the shared backbone for both branches. We utilize SAM2.1 with the Hiera-Large architecture as our foundation model.

Training Protocol. We adopt a two-stage training strategy. **Stage 1**: Pre-train the detector on LVIS base classes [8] for 20 epochs following OVTrack [12], then use this pre-trained detector for DDPL pseudo-label generation. **Stage 2**: Use the high-quality pseudo-labels obtained from Stage 1 to jointly train both branches with $\mathcal{L}_{\text{total}}$ on 10K video frames (2% of MASA's data) for 20 epochs on 4 RTX 3090 GPUs. For more details, such as the parameter setting, training strategies, inference pipeline, etc, please see the **supplementary materials**.

Table 1: Comparison of tracking performance on validation and test sets. We compare the methods on open-vocabulary TAO benchmark [12]. All methods use the same backbone. † represents using the same detector. Baseline results are cited from SLAck [14] under the same evaluation protocol.

Method	Venue (year)	Classes		Novel				Base			
		Base	Novel	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
Validation set											
QDTrack [6]	TPAMI (2023)	✓	✓	22.5	42.7	24.4	0.4	27.1	45.6	24.7	11.0
TETer [11]	ECCV (2022)	✓	✓	25.7	45.9	31.1	0.2	30.3	47.4	31.6	12.1
ByteTrack† [28]	ECCV (2022)	✓	-	22.0	48.2	16.6	1.0	28.2	50.4	18.1	16.0
OC-SORT† [1]	CVPR (2023)	✓	-	23.7	49.6	20.4	1.1	28.9	51.4	19.8	15.4
OVTrack† [12]	CVPR (2023)	✓	-	27.8	48.8	33.6	1.5	35.5	49.3	36.9	20.2
MASA (R50)† [13]	CVPR (2024)	-	-	30.0	54.2	34.6	1.0	36.9	55.1	36.4	19.3
SLAck† [14]	ECCV (2024)	✓	-	31.1	54.3	37.8	1.3	37.2	55.0	37.6	19.1
OVTR [10]	ICLR (2025)	✓	-	31.4	54.4	34.5	5.4	36.6	52.2	37.6	20.1
TRACT [15]	ICCV (2025)	✓	-	31.3	52.7	37.8	3.4	38.5	55.0	39.0	**21.5**
Ours†	-	-	-	**35.7**	**59.0**	**42.4**	**5.6**	**40.1**	**58.1**	**42.3**	19.8
Test set											
QDTrack [6]	TPAMI (2023)	✓	✓	20.2	39.7	20.9	0.2	25.8	43.2	23.5	10.6
TETer [11]	ECCV (2022)	✓	✓	21.7	39.1	25.9	0.0	29.2	44.0	30.4	10.7
OVTrack† [12]	CVPR (2023)	✓	-	24.1	41.8	28.7	1.8	32.6	45.6	35.4	16.9
SLAck† [14]	ECCV (2024)	✓	-	27.1	49.1	30.0	2.0	34.7	52.5	35.6	16.1
OVTR [10]	ICLR (2025)	✓	-	27.1	47.1	32.1	2.1	34.5	51.1	37.5	14.9
TRACT [15]	ICCV (2025)	✓	-	27.3	48.2	30.7	3.1	36.2	52.3	39.1	**17.2**
Ours†	-	-	-	**28.9**	**50.5**	**33.0**	**3.2**	**39.2**	**57.3**	**43.1**	17.1

4 Experiments

4.1 Experimental Setup

Datasets. We conduct comprehensive evaluations on the TAO dataset [5], following the open-vocabulary evaluation protocol established by OVTrack [12]. TAO contains 2,907 videos across 833 object categories with a long-tail distribution. Following the standard protocol, we adopt the same category division as LVIS [8], designating rare categories as novel classes and the remaining as base classes.

Evaluation Metrics. We employ the standard OVMOT evaluation metric, Tracking Every Thing Accuracy (TETA) [11], which provides a comprehensive assessment through three key components: (1) Localization Accuracy (LocA) measuring detection quality, (2) Classification Accuracy (ClsA) evaluating category prediction performance, and (3) Association Accuracy (AssocA) assessing temporal tracking consistency.

4.2 Comparison with State-of-the-Art Methods

We evaluate SOVTrack against existing OVMOT methods and classical MOT approaches on TAO in Table 1.

Substantial Localization Improvements. Our method achieves remarkable localization accuracy gains, demonstrating 58.1% base LocA and 59.0% novel LocA on the validation set. These represent significant improvements over

the strongest competing methods in Table 1. The substantial LocA improvements stem from our Dual-Direction Pseudo Labeling mechanism, which ensures high-quality pseudo-label generation through bidirectional temporal validation. This precise pseudo-label construction enables more accurate object localization, particularly beneficial for novel categories where traditional methods struggle with limited training data.

Superior Association Performance. The most striking improvements emerge in association accuracy (AssocA), which directly reflects tracking quality and temporal consistency. Our method achieves 42.3% base AssocA and 42.4% novel AssocA on the validation set, surpassing SLAck by 4.7% and 4.6% respectively. These exceptional association results validate our core hypothesis that SAM2's rich representations, when properly distilled through our dual-branch architecture, provide superior understanding of object relationships across time. The high association accuracy demonstrates the effectiveness of our Multi-Cue Adversarial Distillation in capturing comprehensive temporal dynamics.

Comprehensive Performance Analysis. SOVTrack establishes new state-of-the-art performance across all evaluation metrics. On the validation set, our method achieves 40.1% TETA for base classes and 35.7% for novel classes, representing substantial improvements of 2.9% and 4.6% over SLAck respectively. The performance gains are even more pronounced on the test set, where we attain 39.2% base TETA and 28.9% novel TETA, demonstrating the robustness and generalizability of our approach across different data distributions. These comprehensive improvements validate the effectiveness of our self-supervised foundation model distillation framework even with minimal video data.

4.3 Ablation Study

We conduct comprehensive ablation studies to analyze the contribution of each component in SOVTrack in Table 2.

DDPL Components. The cycle consistency filtering mechanism demonstrates significant effectiveness in improving tracking quality. Without this filtering, performance drops substantially (2.0% base TETA, 3.5% novel TETA), as unfiltered detections introduce numerous noisy targets that compromise both detection and association accuracy. The degradation is particularly pronounced in AssocA (3.9% base, 4.8% novel), confirming that noise in pseudo-labels directly impacts association learning quality. When using SAM auto-segmentation results as prompts, performance drops by a margin (1.3% base TETA, 3.2% novel TETA) because SAM's segmentation results still contain many noisy segments that lack the precision required for reliable tracking initialization.

Dual-Branch Architecture Components. The dual-branch structure provides substantial improvements over single-branch alternatives. Removing the adversarial branch causes notable performance drops (0.8% base TETA, 2.7% novel TETA), particularly for novel categories where its comprehensive representations prove most valuable. The two branches demonstrate effective complementary capabilities: the contrastive branch excels at base category tracking (+1.3% base TETA), while the adversarial branch shows superior novel

Table 2: Comprehensive ablation study results on the TAO validation set. We systematically analyze the contribution of different components by removing them from our complete SOVTrack framework.

Method Variant	Novel				Base			
	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
DDPL Components								
w/o Cycle Consistency Filtering	32.2	55.4	37.6	3.6	38.1	56.3	38.4	19.5
SAM Auto-segmentation as Prompts	32.5	56.3	37.3	3.8	38.8	56.9	40.2	19.3
Dual-Branch Architecture Components								
w/o contrastive branch	34.4	57.8	40.8	4.5	38.8	58.0	38.9	19.4
w/o adversarial branch	33.0	57.5	37.8	3.7	39.3	57.7	40.9	19.4
w/o localization cue ($\mathbf{f}_{\text{loc}}$)	34.5	57.4	40.7	5.3	39.5	57.3	41.3	**19.9**
w/o association cue ($\mathbf{f}_{\text{assoc}}$)	34.2	58.3	40.6	3.7	38.9	57.7	39.8	19.3
w/o classification cue ($\mathbf{f}_{\text{cls}}$)	33.7	57.6	40.6	2.9	39.2	58.0	40.7	19.0
w/o Adaptive Feature Fusion	33.2	56.9	39.7	3.1	38.8	**58.2**	39.9	18.3
Training Process Components								
w/o Adversarial Distillation	34.2	57.3	40.1	5.1	39.3	57.7	40.7	19.6
w/o $\mathcal{L}_{\text{fusion}}$	34.6	58.1	41.1	4.6	39.7	57.7	41.7	19.6
w/o $\mathcal{L}_{\text{loc}}$	32.8	55.3	40.9	2.1	38.9	55.3	41.6	19.7
SOVTrack (Full)	**35.7**	**59.0**	**42.4**	**5.6**	**40.1**	58.1	**42.3**	19.8

category performance (+2.7% novel TETA). The adaptive feature fusion mechanism further contributes 1.3% base and 2.5% novel TETA gains by intelligently balancing both branches based on tracking difficulty. Within multi-cue fusion, each cue contributes meaningfully: localization cues ($\mathbf{f}_{\text{loc}}$) provide +0.6%/+1.2% (base/novel TETA), association cues ($\mathbf{f}_{\text{assoc}}$) +1.2%/+1.5%, and classification cues ($\mathbf{f}_{\text{cls}}$) +0.9%/+2.0%.

Training Process Components. The localization loss $\mathcal{L}_{\text{loc}}$ contributes significantly (1.2% base TETA, 2.9% novel TETA), with substantial LocA improvements (2.8% base, 3.7% novel), confirming SAM2's high-quality masks provide superior supervision. Adversarial distillation enhances knowledge transfer (0.8% base TETA, 1.5% novel TETA), validating distributional alignment over point-wise L2 regression. The auxiliary fusion loss $\mathcal{L}_{\text{fusion}}$ preserves discriminative properties (0.4% base TETA, 1.1% novel TETA) with notable AssocA gains (0.6% base, 1.3% novel).

4.4 Computational Efficiency Analysis

Despite the additional adversarial branch, SOVTrack maintains computational efficiency. As shown in Table 3, our method achieves 12.1 FPS on a single RTX 3090 GPU, only marginally below MASA's 13.4 FPS under the same detector and backbone. More importantly, DDPL requires only 3.6 GPU-hours for detection-guided VOS processing, compared with MASA's 227.8 GPU-hours for full-image segmentation, yielding a 63× reduction in SAM processing time. Overall, DDPL reduces SAM cost through detector-guided VOS and cycle-closure filtering, while adding no extra SAM calls at inference.

Table 3: Computational efficiency comparison on RTX 3090.

Methods	**TETA**		**FPS**	**SAM Time** (h)
	Base	Novel		
OVTrack	35.5	27.8	1.8	–
MASA	36.9	30.0	13.4	227.8
Ours	**40.1**	**35.7**	12.1	**3.6**

5 Conclusion

We have presented SOVTrack, a novel self-supervised framework for OVMOT that effectively leverages SAM2's powerful representations without requiring manual annotations. Our approach has introduced Dual-Direction Pseudo Labeling for automatic pseudo-label generation, a dual-branch architecture that separates appearance learning from knowledge distillation, and Multi-Cue Adversarial Distillation for capturing SAM2's multi-cue representations. Extensive experiments have demonstrated that our method achieves state-of-the-art performance using only 10K video frames (2% of existing methods' data requirements), with substantial improvements in both base and novel category tracking performance.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62572349, 62402490, and 62422118, in part by the Shenzhen Basic Research Foundation under Grant QNXMC20250701101037019, in part by the Hong Kong Research Grants Council under Grants N_CityU1114/25 and 11219324, and in part by the Hong Kong Innovation and Technology Fund ITS/164/23.

References

1. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9686–9696 (2023)
2. Chen, T., Lu, A., Zhu, L., Ding, C., Yu, C., Ji, D., Li, Z., Sun, L., Mao, P., Zang, Y.: SAM2-Adapter: Evaluating & adapting Segment Anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. arXiv preprint arXiv:2408.04579 (2024)
3. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: SAM-Adapter: Adapting Segment Anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3367–3375 (2023)
4. Cheng, Y., Li, L., Xu, Y., Li, X., Yang, Z., Wang, W., Yang, Y.: Segment and track anything. arXiv preprint arXiv:2305.06558 (2023)

5. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: A large-scale benchmark for tracking any object. In: Computer Vision – ECCV 2020. pp. 436–454. Springer (2020)
6. Fischer, T., Huang, T.E., Pang, J., Qiu, L., Chen, H., Darrell, T., Yu, F.: QDTrack: Quasi-dense similarity learning for appearance-only multiple object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 15380–15393 (2023). https://doi.org/10.1109/TPAMI.2023.3301975
7. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (ICLR) (2022)
8. Gupta, A., Dollár, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5356–5364 (2019)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023)
10. Li, J., Yu, E., Chen, S., Tao, W.: OVTR: End-to-end open-vocabulary multiple object tracking with transformer. In: International Conference on Learning Representations (ICLR) (2025)
11. Li, S., Danelljan, M., Ding, H., Huang, T.E., Yu, F.: Tracking every thing in the wild. In: Computer Vision – ECCV 2022. pp. 498–515. Springer (2022)
12. Li, S., Fischer, T., Ke, L., Ding, H., Danelljan, M., Yu, F.: OVTrack: Open-vocabulary multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5567–5577 (2023)
13. Li, S., Ke, L., Danelljan, M., Piccinelli, L., Segu, M., Van Gool, L., Yu, F.: Matching anything by segmenting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18963–18973 (2024)
14. Li, S., Ke, L., Yang, Y.H., Piccinelli, L., Segù, M., Danelljan, M., Van Gool, L.: SLAck: Semantic, location, and appearance aware open-vocabulary tracking. In: Computer Vision – ECCV 2024. pp. 1–18. Springer (2024)
15. Li, Y., Jiao, Y., Meng, D., Fan, H., Zhang, L.: Attention to trajectory: Trajectory-aware open-vocabulary tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14390–14398 (2025)
16. Peng, Z., Xu, Z., Zeng, Z., Wen, C., Huang, Y., Yang, M., Tang, F., Shen, W.: Understanding fine-tuning CLIP for open-vocabulary semantic segmentation in hyperbolic space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4562–4572 (2025)
17. Pu, X., Jia, H., Zheng, L., Wang, F., Xu, F.: ClassWise-SAM-Adapter: Parameter-efficient fine-tuning adapts Segment Anything to SAR domain for semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 4791–4804 (2025). https://doi.org/10.1109/JSTARS.2025.3532690
18. Qian, Z., Han, R., Hou, J., Song, L., Feng, W.: VOVTrack: Exploring the potentiality in raw videos for open-vocabulary multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7472–7482 (2025)
19. Qian, Z., Han, R., Wang, Z., Hou, J., Feng, W.: COVTrack: Continuous open-vocabulary tracking via adaptive multi-cue fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10054–10063 (2025)

20. Qian, Z., Han, R., Wang, Z., Hou, J., Feng, W.: DOVTrack: Data-efficient open-vocabulary tracking. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 38, pp. 91190–91215 (2026)
21. Qian, Z., Han, R., Wang, Z., Wan, L., Feng, W.: GOVTrack: Towards generative open-vocabulary multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 1872–1882 (2026)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
23. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR) (2025)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 28 (2015)
25. Stojnić, V., Kalantidis, Y., Matas, J., Tolias, G.: LPOSS: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9794–9803 (2025)
26. Wu, X., Zhu, F., Zhao, R., Li, H.: CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7031–7040 (2023)
27. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track Anything: Segment Anything meets videos. arXiv preprint arXiv:2304.11968 (2023)
28. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: Computer Vision – ECCV 2022. pp. 1–21. Springer (2022)
29. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: Computer Vision – ECCV 2022. pp. 350–368. Springer (2022)
30. Zhu, C., Xiao, B., Shi, L., Xu, S., Zheng, X.: Customize Segment Anything Model for multi-modal semantic segmentation with mixture of LoRA experts. arXiv preprint arXiv:2412.04220 (2024)