DIOR: DIstill Observations to Representations for Multi-Object Tracking and Segmentation

Jiarui Cai, Yizhou Wang, Hung-Min Hsu, Haotian Zhang, Jenq-Neng Hwang
University of Washington
Seattle, WA, USA
{jrcai, ywang26, hmhsu, haotiz, hwang}@uw.edu

Abstract

*Multi-object tracking (MOT) has long been a crucial topic in the field of autonomous driving and security monitoring. With the saturation of the bounding-box-based MOT algorithms in recent years, a new task to track objects with instance segmentation, called multi-object tracking and segmentation (MOTS), provides a finer level of scene understanding and introduces potential improvements in tracking accuracy. In this paper, we introduce a video-based MOTS framework, named **DI**still **O**bservations to **R**epresentations (DIOR). A feature distiller is designed to extract and balance the comprehensive object representations: 1) the temporal distiller aggregates context information for consistency of features and smoothness of prediction longitudinally; 2) the spatial distiller on the target of interest within each bounding box removes ambiguity and irrelevance of background in the learned features. The subsequent tracking steps start with Hungarian matching based on feature similarity and masks continuity, which is efficient and straightforward. In addition, we propose short-term retrieval (STR) and long-term re-identification (re-ID) modules to avoid missing associations due to failures in detection or possible occlusion. Our method achieves state-of-the-art performance in both MOTS20 and KITTI-MOTS benchmarks.*

1. Introduction

Multi-object tracking (MOT) has been a hot spot in the computer vision research community due to the increasing demands on various applications, including autonomous driving [12, 45], security monitoring [47, 35], intelligent transportation systems [10, 30, 11] and smart city [29, 9, 41]. In general, MOT targets to associate the objects across time by their locations and appearance features. Therefore, how to extract, fuse and match the features that belong to the same object and distinguish the difference of others become the key challenges.

Current MOT frameworks can be divided into two mainstreams, the tracking-by-detection paradigm and the joint detection and tracking paradigm. In the former one, bounding boxes hypothesis of objects are generated by the video object detector frame-by-frame and followed by a tracklet association stage utilizing the appearance features from a deep feature extractor [37, 5, 34, 3]. In recent years, researchers start to combine the detector with features extractor either by inserting the previous frame's results as the current frame's initial proposals or adding an extra embedding head to the detector [4, 1, 36, 46]. The joint method is neat and efficient, making the MOT possible to be real-time and boosting up the practical applications [16, 22]. However, both of the previous frameworks are based on a similar intuition – extract the object-level features in bounding boxes. When the target boundary is hard to estimate or occlusions exist, the quality of bounding-box features is ambiguous for localization in a single frame, not to mention to be used for matching multiple objects longitudinally. Recently, as the performance of tracking by the bounding boxes is about saturating, researchers try to seek the light at the end of the tunnel by extracting more useful representations and fusing multiple levels of information.

Back to the nature of tracking, human-beings link the objects through different scenes not only rely on the appearance and locations of the objects, but also on the candidates' boundaries, the most distinguishable regions, and surrounding environments. Therefore, by integrating the segmentation into tracking, the multi-object tracking and segmentation (MOTS) [32, 25] pushes the research into a new stage – the feature sources used for tracking algorithms are not only including the object-level information (*i.e.*, categories, bounding box location and size), but also the pixel-level features (*i.e.*, boundary, foreground, and background). Being considered as a new direction of potential improvements, however, pixel-level information is sometimes overabundant. How to integrate both levels becomes a new challenge. Thus, an attention mechanism for extracting, fusing

and balancing the multi-level features is necessary for the MOTS to avoid information redundancy.

In this paper, we propose a multi-level feature embedding and fusion framework for MOTS – **DI**still **O**bservations to **R**epresentations (DIOR), which aims to detect, segment and track objects in video sequences. A feature distiller is designed to extract and balance the comprehensive object representations. By such a customized embedding component, the spatio-temporal information, object-level and pixel-level features can be sufficiently extracted and fused to achieve reliable and precise multi-object tracking in heavy occlusion scenarios. Specifically, the paper yields the following contributions:

- A novel MOTS framework named DIOR, which is a complete workflow for multi-object feature embedding, detection, segmentation and tracking with only monocular videos as input, is introduced. It is a uniform solution for pedestrian tracking in crowded scenes and car tracking for various driving scenarios.
- An observations distiller is designed to transform the raw image sequence into instance-aware embeddings, fusing class, bounding box, mask, key distinguishable regions and appearance features by cross-perspective attention mechanism. With the distiller, the input video clips can be well captured and summarized into feature spaces and fully use in the tracking algorithm.
- A multi-object tracking algorithm is implemented using the jointly learned instance-aware representations along with the temporal consistency of trajectories. The short-term and long-term re-ID steps help to reduce ID switches caused by occlusion.
- Extensive experiments are conducted on two real-world large-scale datasets which consist of pedestrian and vehicle tracking including static and moving cameras. The proposed DIOR achieves the state-of-the-art performance on the MOTS20 benchmark with sMOTSA 69.5% and IDF1 70.3%, and the KITTI-MOTS benchmark with sMOTSA 76.5% (car) and 63.9% (pedestrian).

2. Related Work

Multi-Object Tracking. Tracking multiple objects in the video, *i.e.*, multi-object tracking (MOT), especially persons and vehicles, are extensively studied in recent years. Most prior works [37, 5, 34, 3] perform association based on hypothesis locations from off-the-shelf detectors. These frameworks utilize a separate deep neural network model to extract object appearance features and perform tracking along with the trajectory continuity. The redundant and disjoint use of two parts usually results in degraded performance. Therefore, the recent prevailing stream for MOT is based on joint detection and embedding (JDE) [1, 36, 46], *i.e.*, performing detection and feature learning in the same network, which are more run-time efficient. However, their association is simplified to connection between two adjacent frames [1] or retrieve shortly lost objects based on appearance features [36]. In this paper, our proposed method deploys both short-term re-linking and long-term reID, utilizing the appearance, location and motion featuers.

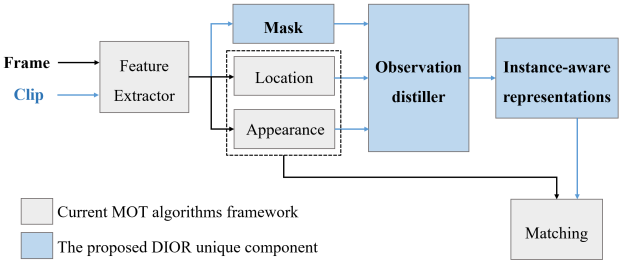


Figure 1. The comparison between the general MOT framework and the proposed DIOR, where the differences are shown in blue. In addition to the use of instance segmentation, we also introduce a distiller module to fuse the bounding-box-level and pixel-level features in both temporal and spatial domains. Our instance-aware features are more informative and less redundant for multi-object tracking and segmentation tasks.

Multi-Object Tracking and Segmentation. The concept and dataset of MOTS was proposed by Voigtlaender *et al.* [32] in 2019, where the instance masks of the MOT17 [22] and KITTI [6] datasets are semi-automatically labeled [21]. The baseline method of MOTS, named Track R-CNN, is a conventional Mask R-CNN [7] with two temporal 3D convolution layers to incorporate the adjacent frames and an additional embedding head to extract instance features for tracking. The embedding head idea is consistent with the aforementioned joint detection and tracking, which is efficient and practical. However, their presented ablation study of tracking performances (in sMOTSA, MOTSA and MOTSP, as formulated in Section 4.2 show only slight increases in the pedestrian category but decreases in those of the car category. Therefore, in this paper, we explore a more effective structure for context aggregation based on attention learning.

As for the KITTI dataset [6], which is mainly for autonomous driving applications. There are several algorithms [20, 43] take multiple sensors as input in addition to monocular images, such as lidar point clouds, stereo images, GPS, and *etc*. The fusion of various data sources enables 3D scene understanding and solves the occlusion issue with more clear clues. However, temporally and spatially synchronized multi-modal data are not commonly unified for general applications and the performance of some of

those sensors are commonly sensitive to object sizes and distances. In this paper, we confine to leverage the use of monocular video only.

Video Object Segmentation and Mask Refinement. Video object segmentation, also known as visual object segmentation (VOS), aims at tracking and segmenting an arbitrary target specified in the first frame throughout the entire video sequence. Same as MOTS, VOS also requires the smoothness of masks longitudinally. However, as only the masks of the first frame are provided, VOS study is dominated by detector's fine-tuning and target re-ID, which is too heavy to be integrated into MOTS workflow directly. In general, VOS studies with temporal aggregation or mask refinement modules, including Youtube-VOS, RVOS and PReMVOS, are correlated with MOTS and our proposal [38, 31]. PReMVOS [21] focuses on mask refinement throughout the video and achieves promising results on DAVIS dataset segmentation challenges [25]. The framework consists of four neural networks for mask generation, optical flow estimation, mask refinement and object re-ID, respectively. However, the pipeline is too large to be run in real-time. Though there are no off-the-shelf models for MOTS in VOS study, the successes of these challenge-winning algorithms inspire us to utilize temporal information to refine single-frame results.

Our proposed DIOR overcomes the issues of the above designs and results in a more flexible and reliable end-to-end single network solution, which also allows effective track recovery after long-time occlusions.

3. The Proposed Method

The workflow of DIOR is shown in Figure 1. Video frames are sent into the temporal feature distiller sequentially in a sliding window (clips), utilizing the adjacent frames to concentrate on the area with objects in the keyframe. It complements the detection difficulty caused by occlusion and ensures the consistency of features longitudinally. The generated clip-level feature for the target object is distilled to emphasize the foreground with spatial attention, eliminating the redundancy of the background. Moreover, this observation embedding not only represents the appearance information but also be trained to minimize intra-instance similarity and enlarge inter-instance diversity. Finally, the features are associated with Hungarian matching [15] efficiently.

3.1. Temporal Feature Distiller

The Temporal Feature Distiller (TFD) is mainly achieved by a temporal attention (TA) module, as shown in Figure 2 (b), which is to smooth the embeddings time-wisely and distill context information to focus on objects on the scene. With features of three consecutive frames x_{t-1}, x_t and x_{t+1} generated by a backbone Ψ_θ, the TA module learns a pixel-wise attention map for each time stamp and uses them to compute weighted fused features. Feature maps of the three frames are concatenated and 3D convolutions are applied to couple them in the time dimension. Then SoftMax is performed over each pixel location to generate the TA maps, which are used to create the TA feature that is the weighted sum of the original backbone features. Formally, the TA maps for each frame τ is represented as

$$\alpha_\tau = \sigma(\mathbb{W}_1[\Psi_\theta(x_\tau)])_\tau, \tag{1}$$

where $\mathbb{W}_1$ means a 3D convolution operation and σ is SoftMax operation. Then the center frame's appearance feature and the weighted TA features are concatenated

$$F^{tem} = \mathbb{W}_2[\Psi_\theta(x_t) \oplus \sum \alpha_\tau \cdot \Psi_\theta(x_\tau)], \tag{2}$$

where $\mathbb{W}_2$ is a 3D convolution and $\oplus$ represents concatenation. The fused features are sent to the following region proposal network (RPN) and prediction heads. The sliding-window input produces results for the center frame.

Since there are no abrupt changes in the background, and object occlusion as well as motion blur are the main causes of mask discontinuity, incorporating the neighboring frames' features could be a useful supplement. Besides, for partially occluded objects, TFD learns to borrow information fore-and-aft, thus increases the recall of detection. Examples of learned TA maps are shown in Figure 3, which shows higher weights are assigned on the frames that objects are less occluded.

3.2. Instance-Aware Representations Learning

In the MOTS scenarios, the crowd is much denser than common object segmentation scenes, such as those in COCO [18]. Therefore, it is likely to include multiple objects within the same bounding box, resulting in erroneous segmentation results. The intuition of the instance-aware features learning comes from the human-beings experience – extracting the distinguishable region at the instance or pixel level for matching candidates. Here, we use the spatial attention (SA) module to highlight the boundary of objects, and then dispatch the foreground (target of interest) and suppress the background, so that more concentrated instance appearance features can be obtained. Details of the prediction heads are shown in Figure 2 (c).

The ROI features for object i, denoted as F_i^{tem}, are pooled and flattened for classification and bounding box regression. Simultaneously, they are passed through an SA module to generate an SA map

$$\beta_i = \delta(\phi_1(F_i^{tem})), \tag{3}$$

where ϕ_1 are 2D convolutional layers shared with the mask head and δ is a pixel-wise Sigmoid operation. The values in

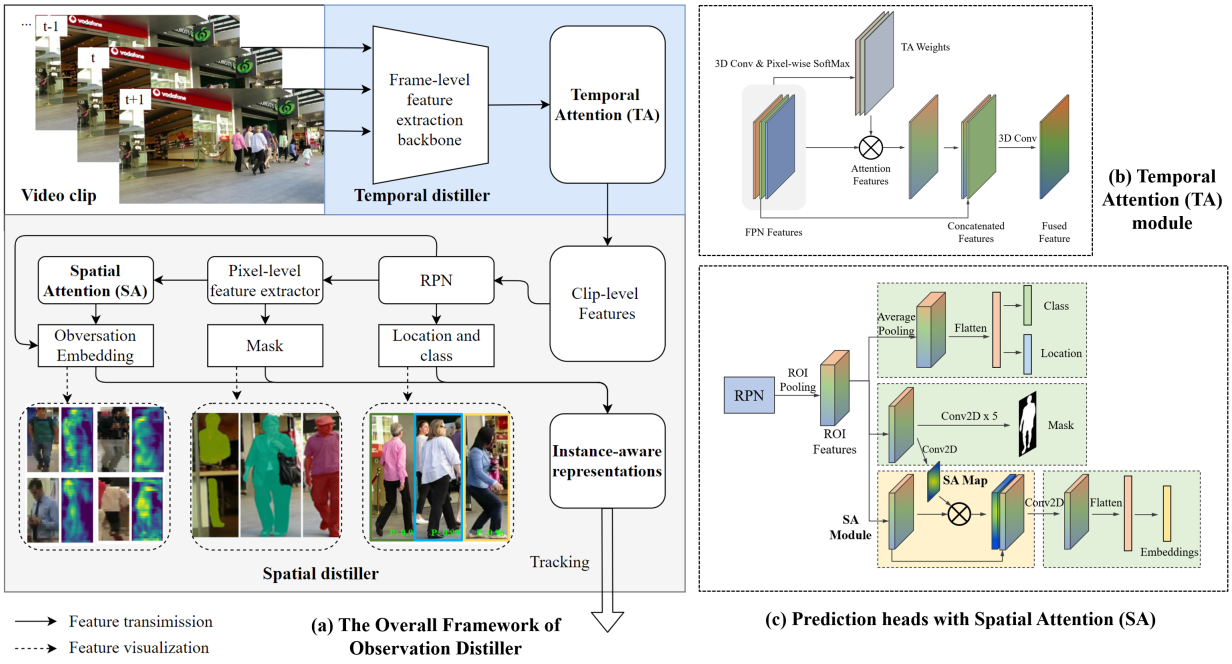


Figure 2. **(a)** Framework for our proposed DIOR. Features are produced by the backbone with shared weights for three consecutive frames, which are aggregated by a temporal attention (TA) sub-network. The proposals are generated by a typical region proposal network (RPN). There are three prediction heads in parallel: a bounding box head for object classification and localization, a mask head for segmentation, and an embedding head for appearance feature extraction. A spatial attention (SA) module, which is inserted between the mask and the embedding head, will heavily weigh on the foreground object to enhance instance-specific appearance features and suppress the noise in the background. **(b)** Detailed structure of the TA module. **(c)** Detailed structure of the prediction heads with SA module.

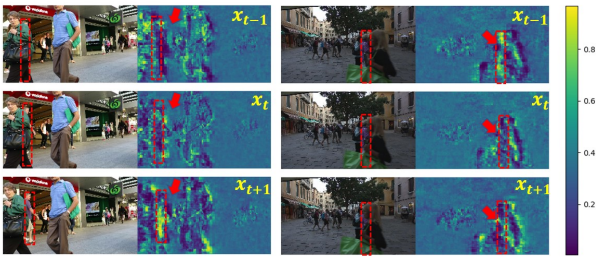


Figure 3. Two examples of TA map visualization, which are best viewed in color. The first column is input images and the second column is TA maps. (Left) The man (labeled in red bounding box) is occluded by the lady in the green shirt in frame $t-1$ and frame t. The feature of frame t is unclear on this bounding box area. Thus higher weights are learned on frame $t+1$, which can supplement the center frame. (Right) The lady on the pink top is gradually occluded and we observe weight decay in the three adjacent frames.

Figure 4. Examples for SA map visualization. For each object, the left figure is the cropped and resized bounding box on the raw image, and on the right is the corresponding SA map. The SA maps highlight the foreground of the proposals and suppress the noisy background, helping to extract more purified and target-specific appearance features.

SA map indicate the probability of objectiveness. Then the SA feature is

$$F^{spa} = \phi_2(F_i^{tem})) \oplus (\beta_i \cdot \phi_2(F_i^{tem}))), \tag{4}$$

where ϕ_2 are convolution layers in the embedding head. Then single-dimensional feature is further extracted by fully connected layers. Figure 4 shows visualizations of SA maps on the validation set, which are rough masks with different weights on specific human body parts through implicit learning.

3.3. Objective Function

The objective function of the DIOR network is

$$L_{total} = L_{bbox} + L_{cls} + L_{mask} + L_{emb}, \tag{5}$$

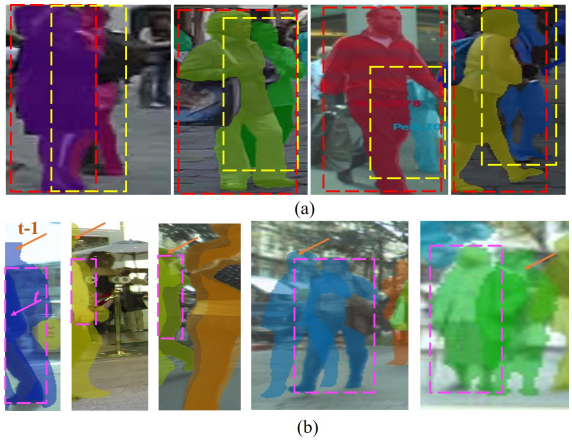


Figure 5. (a) The object in red bounding box and object in yellow bounding box occlude each other. Using bounding box IoU to measure the overlap between the two can cause ambiguity. On the other hand, the overlapping of masks makes better sense. (b) Objects existing in the scene and being gradually occluded have incomplete masks, which change significantly in the heights and widths than those of its previous frame. Here, orange arrow (light color mask) indicates mask in frame $t-1$ and magenta bound box (dark color mask) indicates mask in frame t. To better handle the frequent occlusion as well as leaving and approaching the camera, using minimum mask area between the two is more practical than the union of them.

where L_{bbox} denotes the bounding box location regression loss, which is the smooth L1 loss. L_{cls} is the cross-entropy classification loss. L_{mask} is the binary cross-entropy loss for segmentation. Due to the insufficiency of MOTS training data, the instance embedding is trained with cross-entropy loss instead of triplet loss [26], based on the recognition results of identities of the persons included in the training set. The experiments in [36] also support that cross-entropy loss is more effective and convenient than triplet loss in training object embedding features. Therefore, we also apply a fully-connected layer on the embedding features for object identity classification. L_{emb} is the cross-entropy loss for identity recognition, which helps to generate more distinctive embeddings.

3.4. Tracking with Short-term Retrieval and Long-term Re-ID

The detections, masks and embedding features are first sent to the Hungarian algorithm [15] to match with previously tracked objects of the same object classes (i.e., pedestrians or cars). The predictions of frame t are denoted as $P^{(t)}$, consisting of detections $D^{(t)} = \{d_1^{(t)}, d_2^{(t)}, ..., d_N^{(t)}\}$, masks $M^{(t)} = \{m_1^{(t)}, m_2^{(t)}, ..., m_N^{(t)}\}$, the corresponding embeddings $E^{(t)} = \{e_1^{(t)}, e_2^{(t)}, ..., e_N^{(t)}\}$. $Trk = \{Trk_1, Trk_2, ..., Trk_M\}$ represents a set of established tracks. Considering the visibility of an object may vary due to camera movement, features of each track are represented as an online updated stack of its features of first five frames and the most recent five frames, denoted as Trk_j^e. Similarly, Trk_j^d and Trk_j^m are the bounding boxes and masks of Trk_j. The assignment cost $\mathcal{C}$ between the i-th prediction $p_i^{(t)} = \{d_i^{(t)}, m_i^{(t)}, e_i^{(t)}\}$ and j-th track Trk_j is computed by

$$\mathcal{C} = 2 - IoM(m_i^{(t)}, Trk_j^{m^{t-1}}) - \mathcal{D}(\max(e_i^{(t)}, Trk_j^e)), \tag{6}$$

where the $Trk_j^{m^{t-1}}$ is the Trk_j's mask in frame F_{t-1}. $\mathcal{D}(\max(\cdot))$ denotes the maximum of cosine similarity from pair-wisely comparison. Here, the mask IoM (Intersection-over-Min) is defined as

$$IoM(m_a, m_b) = \frac{\mathcal{I}(m_a, m_b)}{\min(\mathcal{S}(m_a), \mathcal{S}(m_b))}, \tag{7}$$

where $\mathcal{S}$ represents the area and $\mathcal{I}$ is the intersection. Using the intersection of masks avoids falsely producing high Intersection-over-Union (IoU) value for two objects that occlude each other, as the example shown in Figure 5 (a). In addition, exiting objects have incomplete masks, which are incomplete in shape compared to their former frames, as the case in Figure 5 (b). Therefore, the minimum area of two masks is used in the denominator of IoM.

Matched tracks are updated with new detections while the unassigned ones are sent into a short-term retrieval module to be matched with the live tracks that without a detection in frame x_{t-1}. Tracklets will be marked as terminated if there is no alignment for the most recent N_1 frames.

In offline applications, re-ID could reduce identity switch (IDS) by reconnecting broken-up tracks. Here, the long-term occlusions are recovered by feature-based re-ID. In this stage, two tracklets Trk_p and Trk_q without overlapped frames in time (assuming Trk_p is earlier than Trk_q), within N_2 frames apart, and with feature similarity higher than θ_1, are considered as possible matched pairs.

4. Experiments

4.1. Dataset and Implementation Details

The datasets for evaluation are MOTS20 and KITTI-MOTS datasets [32]. MOTS20 has 8 sequences for pedestrian tracking only, and is evenly split for training and testing. The videos are captured both indoor and outdoor by static, hand-held, stroller-mounted and car-mounted cameras. In the testing set, the resolution varies from 640×480 to 1920×1080 with an average density of 10.6 targets per frame. The ground truth of MOTS20 dataset is for monocular images only. KITTI-MOTS is a dataset for the autonomous driving scene. It consists of 21 training sequences and 28 testing sequences from car-mounted cameras, covering the street view, high-way and pavement view. The target categories are both cars and pedestrians.

Figure 6. Qualitative results of the proposed method. From left to right: KITTI-MOTS-0027 (mix video); KITTI-MOTS-0000 (vehicle-centric video); MOTS20-0007 (hand-held camera, outdoor, high density); MOTS20-0012 (hand-held camera, indoor).

The proposed network uses ResNet50 [8] as the backbone, which is pretrained on the COCO dataset [18] and fine-tuned on the MOTS20 and KITTI-MOTS dataset. Noted that no other training datasets are included. The dimension of the output feature is 1024. Short-term memory interval to determine the state of a track is $N_1 = 0.2$ second, while the long-term interval for re-ID is $N_2 = 1$ second. Furthermore, to fairly assign hyper-parameters and conduct ablation studies, MOTS20-0002, MOTS20-0005 and MOTS20-0009 are selected as the training set, and MOTS20-0011 is the validation sequence. Results are reported based on our implementations using Pytorch [23] and mmdetection [2] using two Nvidia Titan Xp GPUs on a Linux Ubuntu 18.04 system.

4.2. Evaluation Metrics

Metrics for MOTS [32] is an extension of the CLEAR MOT, established in [22]. Here we briefly introduce several necessarily defined evaluation metrics for MOTS. The correspondence c of ground truth masks m and hypotheses h are established based on mask IoU with threshold 0.5 [13]

$$c(h) = \begin{cases} \arg\max\limits_{m \in M} \text{IoU}(h, m), & \text{if } \max\limits_{m \in M} \text{IoU}(h, m) > 0.5, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{8}$$

Based on the correctness of the masks correspondence, there are sets of true positive (TP), false positive (FP) and false negative (FN). Additionally, soft TP is defined as

$$\widetilde{\text{TP}} = \sum_{h \in \text{TP}} \text{IoU}(h, c(h)) \tag{9}$$

which evaluates the accumulated quality of the segmentation rather than a hard counting with the threshold. ID switches (IDS) is defined as the set of ground truth masks whose predecessor was tracked with a different id. The MOTS accuracy (MOTSA), MOTS precision (MOTSP) and soft MOTSA (sMOTSA) are separately formulated as

$$\begin{aligned} \text{MOTSA} &= 1 - \frac{|\text{FN}| + |\text{FP}| + |\text{IDS}|}{|\text{M}|} \\ \text{MOTSP} &= \frac{\widetilde{\text{TP}}}{\text{TP}}, \ \text{sMOTSA} = \frac{\widetilde{\text{TP}} - |\text{FP}| - |\text{IDS}|}{|\text{M}|}. \end{aligned} \tag{10}$$

4.3. Performance

The performance of DIOR in the MOTS20 benchmark is shown in Table 1. By the date of submission of this paper, there are very few published competitors. Comparing to the baseline TrackRCNN [32] method, we significantly improve the recall by 25.2%, and thus increase the number of mostly tracked (MT) and reduce the mostly lost (ML) to a large extent. TrackRCNN uses a 3D convolution layer to extract context information; differently, we design the attention mechanism on spatial and temporal domains. The result proves the effectiveness of our future distillation.

The performance of DIOR in the KITTI-MOTS is shown in Table 3. PointTrack [39] is one of the leading algorithms in the benchmark. It regards the object's mask and its surrounding environment as two sets of 2D point clouds and learns the foreground and background features separately. Nevertheless, this is built upon an accurate initialization of segmentation, achieved by a pretrained network for optical flow estimation, which is unreliable for unsteady moving cameras (e.g., hand-held or stroller-mounted, as in the MOTS dataset), resulting in degraded performance. ReMOTS [40] proposed an intra-frame self-supervised triplet construction network to learn object features for both training and testing set for re-ID. GMPHD-SAF [27] uses Gaussian mixture probability hypothesis density (GMPHD) filter and a simple affinity fusion (SAF) model. Though computationally simple, the GMPHD-SAF is tracking-only method, relying on well-preprocessed detections and masks. The ReMOTS is a refinement of other methods' tracking results.

Tracker	sMOTSA↑	IDF1↑	MOTSA↑	MOTSP↑	MT↑	ML↓	Recall↑	Precision↑	IDS↓
GMPHD_MAF	69.0	65.6	82.9	84.2	249	11	87.7	96.7	566
ZXPointTrack	62.3	42.9	76.8	82.3	186	41	81.4	96.5	541
SORTS_ReID	55.8	65.8	69.1	81.9	107	52	73.4	95.7	541
TrackRCNN [32]	40.6	42.3	55.2	76.1	127	71	60.8	94.0	567
ReMOTS [40]	70.4	75.0	84.4	84.0	248	9	87.6	97.2	231
DIOR (Ours)	69.5	70.3	83.3	84.2	253	9	87.1	97.2	421

Table 1. Performance on MOTS20 dataset. The results are from `https://motchallenge.net/results/MOTS/`. We submit under name COSTA_st.

Method	KITTI-MOTS		MOTS	Total
	Car	Ped		
MCFPA [28]	77.0	67.2	66.1	69.1
TPM-MOTS [42]	75.8	67.3	66.6	69.1
ReMOTS [40]	72.6	64.6	67.9	68.3
GMPHD_SAF[27]	76.2	64.3	64.3	67.3
Lif_TS	77.5	55.8	65.3	66.0
SRF [17]	71.4	60.9	60.0	63.1
KQD	74.4	61.8	57.3	62.7
USN	72.1	59.3	59.5	62.6
YLC	62.3	57.2	59.1	59.4
SI	68.5	55.5	56.2	59.1
FK [14]	64.1	54.5	54.3	56.8
DIOR (Ours)	76.4	64.0	69.4	69.8

Table 2. Track 3 (tracking-only track with given detection and segmentation) competition results of the BMTT Challenge in CVPR 2020, evaluated by sMOTSA.Information of detailed leader board and teams is available at the challenge host site `https://motchallenge.net/workshops/bmtt2020`

Therefore, our method yields higher integrity and practicability. We are the 2^{nd} place for both car and pedestrian categories.

As a supplement, tracking-only performance is reported in Table 4.2 based on the BMTT Challenge in CVF/IEEE CVPR 2020 workshop (tracking-only track with public detections on both MOTS20 and KITTI-MOTS datasets). The pre-computed detections are generated from Mask R-CNN X152 [7] and refined by the refinement net [19]. MCFPA [28] is based on Min-Cost network Flow (MCF) [44] optimization, then the post association (PA) of tracklets is performed by a single object tracker SPM [33]. TPM-MOTS adjusts TMP [24] trackor with mask input. Processing time and offline requirements are the major drawbacks of these methods. Besides, all of them are following tracking-by-detection scheme, the performance highly relies on detection accuracy. The proposed method could perform detection and feature extraction jointly, which is more robust and efficient.

4.4. Ablation Studies

How does SA module improve tracking performance? As mentioned in Section 3.2, the proposed SA module in spatial distiller manages to suppress the noisy background information, resulting in better feature representations for reliable multi-object tracking. Some visualization results of the SA module is shown in Figure 4. The quantitative results of the contributions for TA and SA to the tracking performance are shown in Table 4. It shows that the SA module can improve sMOTSA by around 20%, which shows our high-quality instance-aware features.

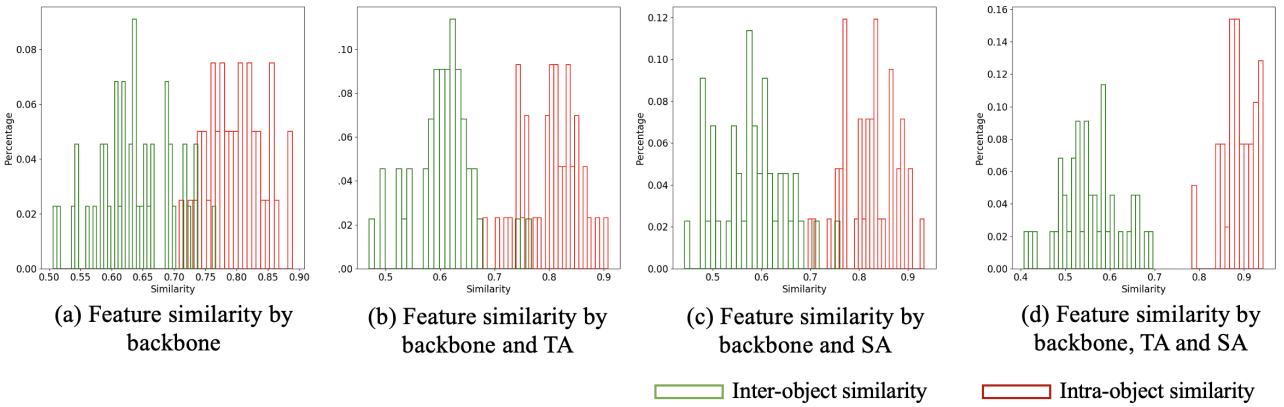
Figure 7 shows the comparison of feature similarity between TA only model and TA+SA model in histogram. Here all detections are assigned to a ground truth bounding box based on their degrees of overlap. Thus, each detection will be assigned an ID if the IoU is above threshold 0.75. Then the inter-object and intra-object feature cosine similarity are calculated pair-wisely. The histogram shows the average similarity and the corresponding number of tracks. It is observed that, with the SA model, object features are more distinctive and separable.

How does TA module improve detection performance? The proposed key component of temporal distiller is the TA module, which can efficiently merge the information from the input clip (mentioned in Section 3.1). Some examples for qualitative visualization of the TA module are shown in Figure 3. From Table 5, we can observe that TA module improves object detection mAP by about 2%. Besides, given the limited improvement in mAP (+0.001) and mAR (+0), we can also find that SA module can hardly help object detection. With SA module, there is a trade-off between detection and embedding features, which means that to obtain better features, the detection performance might degrade.

Short-time Retrieval and Re-ID Modules. The effect of the short-term retrieval and long-term re-ID modules are shown in Table 6. Since they share the same prediction from the DIOR network, the sMOTSA, MOTSA and MOTP are close. However, the number of IDS is significantly reduced with these two modules.

	Tracker	sMOTSA↑	MOTSA↑	MOTSP↑	MOTSAL↑	MODSA↑	MODSP↑	MT↑	ML↓	IDS↓
Car	PointTrack [39]	78.5	90.9	87.1	91.8	91.8	89.7	90.8	0.6	346
	ReMOTS [40]	75.9	86.7	88.2	88.7	88.7	90.7	84.5	0.6	716
	GMPHD-SAF [27]	75.4	86.7	87.5	88.2	88.2	90.1	82.0	0.6	549
	TrackRCNN [32]	67.0	79.6	85.1	81.5	81.5	88.3	74.9	2.3	692
	DIOR (Ours)	76.5	87.4	88.1	89.9	89.9	90.5	84.7	1.1	649
Pedestrian	ReMOTS [40]	66.0	81.3	82.0	83.2	83.2	94.0	62.6	5.6	391
	GMPHD-SAF [27]	62.8	78.2	81.6	80.4	80.5	93.7	59.3	4.8	474
	PointTrack [39]	61.5	76.5	81.0	77.4	77.4	93.8	48.9	9.3	176
	TrackRCNN [32]	47.3	66.1	74.6	68.4	68.4	91.8	45.6	13.3	481
	DIOR (Ours)	63.9	80.3	81.5	83.3	83.3	93.6	73.0	2.2	611

Table 3. Performance on KITTI-MOTS dataset.



(a) Feature similarity by backbone

(b) Feature similarity by backbone and TA

(c) Feature similarity by backbone and SA

(d) Feature similarity by backbone, TA and SA

Figure 7. Inter-object similarity is shown in green, while intra-object similarity is shown in red. x-axis is the average cosine similarity, y-axis is the proportion of tracks. Best viewed in color.

Model	sMOTSA	MOTSA	MOTP	MODSA
base	42.1	50.9	86.9	61.8
+TA	43.6	51.6	**87.4**	61.7
+SA	62.7	72.8	86.7	73.3
+TA+SA	**65.4**	**76.1**	86.5	**76.9**

Table 4. Ablation study of TA and SA modules on tracking. The performance is based on the validation sequence MOTS20-0011.

Model	mAP @[IoU=0.5:0.95]	mAR @[IoU=0.5:0.95]
base	0.569	0.597
+SA	0.570	0.597
+TA	**0.586**	**0.612**
+TA+SA	0.578	0.602

Table 5. Ablation study of TA and SA modules on detection. The performance is based on the validation sequence MOTS20-0011. mAP means the mean average precision at IoU threshold from 0.5 to 0.95, with step of 0.05. mAR is the mean average recall.

Model	sMOTSA	MOTSA	MOTP	IDS
Hungarian only	64.6	75.7	**86.7**	54
+STR	64.7	75.4	**86.7**	48
+STR+re-ID	**65.4**	**76.1**	86.5	**21**

Table 6. Ablation study of Short-term Retrieval (STR) and Long-term re-ID modules. The tracking performance on validation sequence MOTS20-0011. Best results are marked in **bold**.

5. Conclusion

In this paper, we introduce DIOR: distill observations to representations, which is a complete framework for multi-object joint detection, segmentation and tracking with only monocular videos as input. In DIOR, the temporal distiller module incorporates context features to compensate for mask discontinuity caused by occlusion or motion blur and ensure longitudinal consistency. Besides, the spatial distiller module is designed to highlight the target of interest and suppress the redundant background. The distinctive instance-aware representations significantly benefit the object association. Our system achieves leading performance on MOTS benchmarks.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the*

IEEE international conference on computer vision, pages 941–951, 2019.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6172–6181, 2019.

[4] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3038–3046, 2017.

[5] Weitao Feng, Zhihao Hu, Wei Wu, Junjie Yan, and Wanli Ouyang. Multi-object tracking with multiple cues and switcher-aware classification. *arXiv preprint arXiv:1901.06129*, 2019.

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] 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.

[9] Hung-Min Hsu, Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang, and Kwang-Ju Kim. Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model. *IEEE Transactions on Image Processing*, 30:5198–5210, 2021.

[10] Hung-Min Hsu, Tsung-Wei Huang, Gaoang Wang, Jiarui Cai, Zhichao Lei, and Jenq-Neng Hwang. Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In *CVPR Workshops*, pages 416–424, 2019.

[11] Hung-Min Hsu, Yizhou Wang, and Jenq-Neng Hwang. Traffic-aware multi-camera tracking of vehicles based on reid and camera link model. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 964–972, 2020.

[12] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *CVPR Workshops*, volume 2, 2019.

[13] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[14] Franz Koeferl, Johannes Link, and Bjoern Eskofier. Application of sort on multi-object tracking and segmentation.

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[16] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[17] Jimi Lee, Sangwon Kim, and Byoung Chul Ko. Fast multiple object tracking using siamese random forest without online tracker updating. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. BMTT Workshop (CVPRW)*, pages 1–4, 2020.

[18] 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.

[19] B Leibe J Luiten, P Voigtlaender, and B Leibe. Premvos: Proposal-generation. *Refinement and Merging for the DAVIS Challenge on Video Object Segmentation*, 2018.

[20] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2):1803–1810, 2020.

[21] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. *CoRR*, abs/1807.09190, 2018.

[22] 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.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[24] Jinlong Peng, Tao Wang, Weiyao Lin, Jian Wang, John See, Shilei Wen, and Erui Ding. Tpm: Multiple object tracking with tracklet-plane matching. *Pattern Recognition*, page 107480, 2020.

[25] 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.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[27] Young-min Song and Moongu Jeon. Online multi-object tracking and segmentation with gmphd filter and simple affinity fusion. *arXiv preprint arXiv:2009.00100*, 2020.

[28] Jiasheng Tang, Xiong Xiong, Chenwei Xie, Yanhao Zhang, Pichao Wang, Fan Wang, Fei Du, Liang Han, Yun Zheng, Pan Pan, et al. Min-cost network flow and trajectory fix for multiple objects tracking.

[29] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019.

[30] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 108–115, 2018.

[31] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5277–5286, 2019.

[32] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019.

[33] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Spm-tracker: Series-parallel matching for real-time visual object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3643–3652, 2019.

[34] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 482–490, 2019.

[35] Gaoang Wang, Xinyu Yuan, Hung-Min Hsu, and Jenq-Neng Hwang. Anomaly candidate identification and starting time estimation of vehicles from traffic videos.

[36] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2019.

[37] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[38] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[39] Zhenbo Xu, Wei Zhang, Xiao Tan, Wei Yang, Huan Huang, Shilei Wen, Errui Ding, and Liusheng Huang. Segment as points for efficient online multi-object tracking and segmentation. In *European Conference on Computer Vision*, pages 264–281. Springer, 2020.

[40] Fan Yang, Xin Chang, Chenyu Dang, Ziqiang Zheng, Sakriani Sakti, Satoshi Nakamura, and Yang Wu. Remots: Self-supervised refining multi-object tracking and segmentation, 2020.

[41] Hao Frank Yang. *Novel Traffic Sensing Using Multi-Camera Car Tracking and Re-Identification (MCCTRI)*. PhD thesis, 2020.

[42] Yuchen Yuan, Xiangbo Su, Wei Zhang, Tao Wang, Wei Shi, Zhenbo Xu, Mian Peng, Xiao Tan, Weiyao Lin, Hongwu Zhang, et al. Re-identification and tracklet-plane matching for multi-object tracking and segmentation.

[43] Haotian Zhang, Yizhou Wang, Jiarui Cai, Hung-Min Hsu, Haorui Ji, and Jenq-Neng Hwang. Lifts: Lidar and monocular image fusion for multi-object tracking and segmentation.

[44] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[45] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2365–2374, 2019.

[46] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020.

[47] Jianfei Zhao, Zitong Yi, Siyang Pan, Yanyun Zhao, Zhicheng Zhao, Fei Su, and Bojin Zhuang. Unsupervised traffic anomaly detection using trajectories. In *CVPR Workshops*, pages 133–140, 2019.