

CML-MOTS: Collaborative Multi-task Learning for Multi-Object Tracking and Segmentation

Yiming Cui, Cheng Han, Dongfang Liu

Abstract—The advancement of computer vision has pushed visual analysis tasks from still images to the video domain. In recent years, video instance segmentation, which aims to track and segment multiple objects in video frames, has drawn much attention for its potential applications in various emerging areas such as autonomous driving, intelligent transportation, and smart retail. In this paper, we propose an effective framework for instance-level visual analysis on video frames, which can simultaneously conduct object detection, instance segmentation, and multi-object tracking. The core idea of our method is collaborative multi-task learning which is achieved by a novel structure, named associative connections among detection, segmentation, and tracking task heads in an end-to-end learnable CNN. These additional connections allow information propagation across multiple related tasks, to benefit these tasks simultaneously. We evaluate the proposed method extensively on KITTI MOTS and MOTS Challenge datasets and obtain quite encouraging results.

I. INTRODUCTION

In the past decade, the computer vision community has achieved significant progress in many tasks with the development of deep learning [1]–[4]. Among various visual tasks, instance segmentation [5] has drawn wide attention due to its importance in many emerging applications, such as autonomous driving [6]–[11], augmented reality [12], [13], and video captioning [14], [15]. Technically, it is quite challenging as it is a compound task consisting of both object detection and segmentation, each of which is a difficult task and has been studied for a long time.

Compared to instance segmentation on images, multi-object tracking, and segmentation is much more challenging because it not only needs to perform instance-level segmentation on individual frames but also has to depict the coherence of each instance in consecutive video frames [16]. Due to these challenges, multi-object tracking and segmentation have received much attention in recent years [17]–[22]. In general, multi-object tracking and segmentation contain object detection, segmentation, and tracking simultaneously in consecutive video frames. Compared to video object segmentation [23] that deals with object segmentation and tracking in videos, multi-object tracking and segmentation require additional object detection. It also has to generate object masks compared to video object detection which contains both object detection and tracking in videos [24]–[29].

Currently, the state-of-the-art methods are mainly based on Mask R-CNN [5] while adding tracking sub-network. The Mask R-CNN family [19], [22] has demonstrated an appealing performance on this task. However, these methods still

have several drawbacks. For instance, TrackR-CNN [19] uses proposal-based ROI features, which are not fine enough for mask and tracking heads to produce accurate predictions. In detail, the current methods use the coarse proposal-based ROI features directly, which is not enough. Instead, our method processes the coarse proposal-based ROI features first and then uses the refined version for the downstream tasks. In addition, using proposal-based features not only has high computational complexity but also makes it prone to incorrect or redundant predictions. What is more, TrackR-CNN models object movements and scene consistency among video frames by using a 3D convolutional operation, which is also parameter-heavy and computationally expensive. Although PointTrack [22] achieves a significant improvement in segmentation results by proposing a more powerful mask head, it is a multi-step architecture and so cannot be jointly optimized. In PointTrack, detection, and segmentation operations are applied to the input first. This partial model is optimized first without the tracking task. Then the detected and segmented objects are transformed into point cloud formats for the tracking task, which is optimized. In general, the whole model contains multiple steps and is not optimized jointly or end-to-end trainable. Moreover, its point-cloud strategy requires additional post-processing. Meanwhile, although these methods are trained by multiple learning objectives corresponding to different tasks, the relations among these tasks have not been explored.

Intuitively, object detection could benefit instance segmentation, and good object masks are also helpful for multi-object tracking. Inspired by this intuition, we propose a novel idea of collaborative multi-task learning for multi-object tracking and segmentation. Associative connections are added among different task heads to enable information propagation across them. Although our method is also based on Mask R-CNN [5], it fundamentally differs from other Mask R-CNN variants [19], [22] as our method exploits associative connections to facilitate accurate information propagation across the detection, segmentation, and tracking heads to benefit individual tasks. With associative connections, our segmentation and tracking can perform predictions on refined ROI features instead of using features pooled from the very coarse region proposals. Extensive experimental results demonstrate significant improvements over the existing state-of-the-art methods for multi-object tracking and segmentation on two benchmarks (KITTI MOTS [19] and MOTS Challenge [30]). The principal contributions of this work can be summarized as follows,

- We propose a novel idea of collaborative multi-task

Yiming Cui is with the University of Florida, Gainesville, FL 32611, USA.

Cheng Han and Dongfang Liu are with the Rochester Institute of Technology, Rochester, New York, USA.

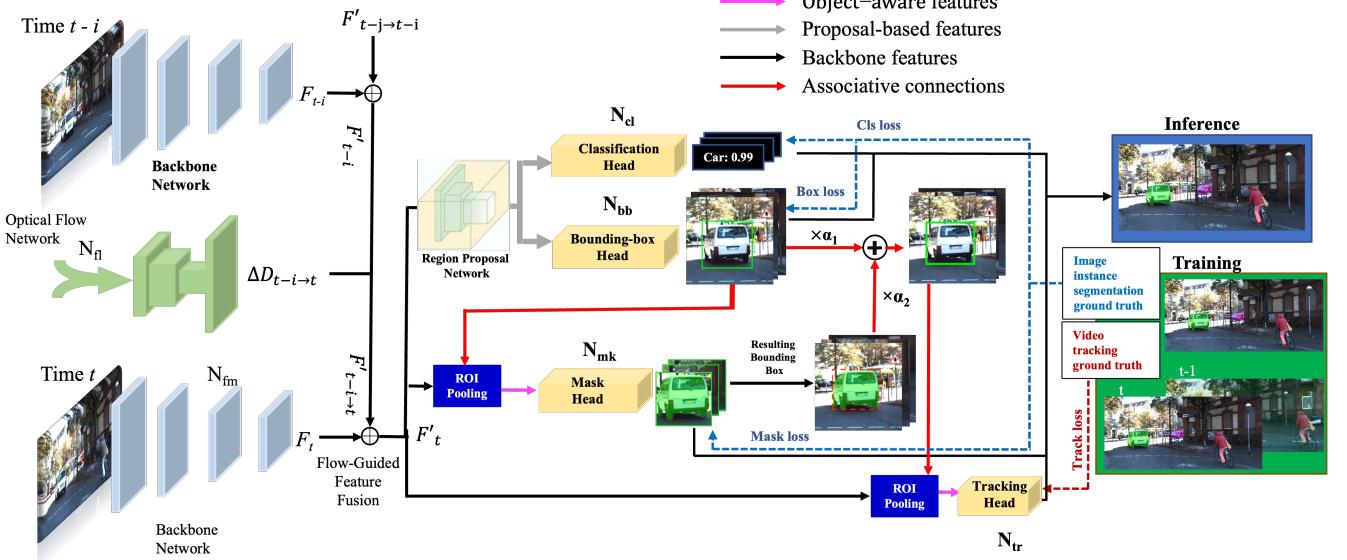


Fig. 1. Illustration of the proposed framework. It adds associative connections among different task heads in TrackR-CNN [19] to enable collaboration among multiple tasks by facilitating information propagation across them. Particularly, the bounding boxes outputted by the detection head are linked to the mask head to extract object-aware features for instance segmentation. The outputs of the mask head are fused with the outputs of the detection head to obtain more reliable bounding boxes, which are then inputted into the tracking head for better tracking ability. Besides, a lightweight optical flow network is used to model the object movements across successive frames, which is used for aligning object features among these frames to obtain more powerful feature maps in the currently processed frame.

learning for multiple object tracking and segmentation in videos.

- We design associative connections among different tasks to enable information flow through different task heads, to simultaneously learn from multiple tasks and consider their intrinsic relations in the meantime.
- Compared to existing methods, our method achieves better results on KITTI MOTS and MOTS Challenge benchmarks, especially on the consistency of multi-object tracking.

II. RELATED WORK

Visual tasks in the video domain have been under-explored in the literature compared to image-level tasks. Particularly, video instance segmentation which includes detection, segmentation, and tracking is largely ignored due to its extreme difficulties. This section offers a review of recent works about several related tasks to video instance segmentation.

A. Image Object Detection

State-of-the-art object detection methods [25], [31]–[37] are generally based on deep CNNs for feature extraction and a shallow detection structure for detection prediction, including classification and bounding box regression. R-CNN [38] proposed a multi-stage pipeline to classify region proposals at different semantic levels for object detection. To speed up, Fast R-CNN [39] and Libra R-CNN [40] used ROI pooling on the feature maps which are shared on the entire image. As a representative work of the multi-stage detection family, Faster R-CNN [31] introduced a Region Proposal Network (RPN) to generate region proposals and then the proposal-based features are shared between classification and bounding box regression heads. R-FCN [41] replaced ROI pooling

with position-sensitivity ROI pooling to further improve the recognition accuracy while still facilitating feature sharing.

Traditional two-stage object detectors, exemplified by the R-CNN family [31], [39], [42], rely on a plethora of predefined anchor boxes to designate initial object candidate locations. In contrast, one-stage methods [43], [44] were introduced to enhance the efficiency and inference speed of object detectors by forgoing the use of region proposals. Recently, query-based approaches [45]–[48] have emerged, replacing anchor boxes and region proposals with learned proposals or queries. DETR [45] adapts an encoder-decoder architecture based on transformers [49] to generate a sequence of prediction outputs. It introduces a set loss function to facilitate bipartite matching between predicted and ground-truth objects. Deformable-DETR [46] enhances the convergence of DETR by refining feature spatial resolutions. Sparse R-CNN [48] employs a fixed sparse set of learned object proposals to classify and localize objects in the image. It utilizes dynamic heads to generate final predictions directly, eliminating the need for post-processing techniques like non-maximum suppression.

B. Image Instance Segmentation

Instance segmentation not only predicts semantic classes on each pixel but also groups pixels into different object instances. Due to the effectiveness of R-CNN [38], many instance segmentation methods perform mask prediction on top of region proposals. Some early methods are based on the bottom-up segment strategy. For instance, the DeepMask family [17], [50] learns to segment proposal candidates first and then classify them using Fast R-CNN. These methods have segmentation precede detection, which is slow and inaccurate.

Another strategy for instance segmentation is based on a parallel prediction of masks and class labels [5]. Li et

al. [51] proposed the fully convolutional instance segmentation by combining the segmentation proposal system in [50] and object detection system in R-FCN [41]. Basically, [50], [51] all use a set of position-sensitive output channels to simultaneously predict object classes, bounding boxes, and masks, thus being fast and fundamentally similar. However, they struggle to deal with occlusion or truncation instances as they tend to create spurious edges in those cases.

Besides two-stage methods, one-stage instance segmentation frameworks like YOLACT [52], [53], SipMask [54], and SOLO [55], [56] have been introduced to strike a balance between inference speed and accuracy. Recently, QueryInst [57] extended the query-based object detection method Sparse R-CNN [48] to the instance segmentation task by incorporating a dynamic mask head and parallel supervision. Nevertheless, all the two-stage and query-based methods mentioned earlier employ a fixed number of proposals, which may not be adaptable to images with varying objects or devices with different computational resource constraints.

C. Object Tracking

Tracking-by-detection is a popular strategy for multi-object tracking [58]–[60]. A common practice is to associate the tracks with detection based on their confidences [60]. To improve the tracking accuracy, Sun et al. [61] exploited multiple detectors by considering outputs from multiple over-detected detectors. However, this method enhances tracking performance at the cost of high computational complexity. More recently, Kim et al. [62] proposed to use a single object tracker based on a binary classifier for online multi-object tracking. Their proposed architecture has shared features for tracking and classification to speed up the process. Even though, [62] is still computationally expensive for real-time applications.

Many previous methods deal with tracking tasks as a global optimization problem [17], [63]. This kind of method formulates temporal information from nearby frames to reduce noisy detection and handle ambiguities for object association. The principal strategy is to re-identify objects using the embedding vectors since each association vector represents the identity of an object [19], [63]. Inspired by the aforementioned methods, our work also leverages deeply learned ReID features. With the proposed associative connections, our work has better object-aware features to obtain improved identification performance.

D. Video Object Detection

Video object detection involves the identification and localization of objects of interest in each frame, even in the presence of potential degradation due to rapid motion. Present approaches [27]–[29], [64]–[72] typically extend image-based object detectors into the realm of videos. These models fall into two categories: Post-processing-based and feature-aggregation-based.

Post-processing-based models extend image object detectors to video by linking prediction results across frames based on temporal relationships [64]–[66]. Examples include T-CNN [64] and Seq-NMS [65]. T-CNN employs a CNN-based pipeline with straightforward object tracking for video object detection. Seq-NMS associates prediction results from each

frame using the IOU threshold. While these models outperform single-image object detectors, they heavily rely on individual frame detections and lack joint optimization. If single-frame outputs are erroneous, the post-processing pipeline cannot rectify them, resulting in suboptimal performance. Furthermore, these models tend to be slower as they process each frame independently before post-processing.

In contrast, feature-aggregation-based and Transformer-based models can aggregate information across frames and jointly optimize predictions, yielding improved performance. These models effectively utilize temporal and spatial cues to track and detect objects across consecutive frames, making them better suited for video object detection tasks. While post-processing-based models offer some improvement over single-image object detectors, they are constrained by their reliance on individual frame results and slower inference. Feature-aggregation-based and Transformer-based models present a more promising approach, leveraging data from multiple frames and optimizing jointly.

Feature-aggregation-based models enhance current frame representations by incorporating features from adjacent frames, assuming they can mitigate feature degradation. Several models embody this concept. For instance, FGFA [28] employs estimated optical flow to fuse neighboring features, while MANet combines pixel-level and instance-level object features. Conversely, SELSA [67] calibrates features based on semantic similarity rather than temporal relations. MEGA [70] integrates local and global temporal information to enhance performance. Although these models surpass post-processing-based ones in performance, they typically demand more computational resources, resulting in slower inference speeds. In summary, while feature-aggregation-based models have enhanced video object detection, they often come at the expense of slower inference. Transformer-based models offer a promising solution to this issue by efficiently fusing information across frames, suggesting further advancements in video object detection tasks.

E. Video Instance Segmentation

Recent endeavors on video instance segmentation [17], [19], [22] are intuitive extensions of Mask R-CNN [5] while considering additional motion cues [73] or temporal consistency [74]–[76] in videos. There are three main categories of existing methods for Visual Instance Segmentation (VIS): two-stage, one-stage, and transformer-based. Two-stage approaches [16] build upon the Mask R-CNN family [5], [31], incorporating an additional tracking branch for object association. One-stage methods [20], [77] employ anchor-free detectors [78], often utilizing linear mask basis combination [52] or conditional mask prediction generation [79]. Transformer-based models [80]–[83] introduce innovative adaptations of the transformer architecture for VIS tasks. VisTr [82] pioneers the application of transformers in VIS, and IFC [84] enhances efficiency through the use of memory tokens. Seqformer [85] introduces frame query decomposition, while Mask2Former [80] incorporates masked attention. VMT [86] extends the Mask Transfiner [87] to video for high-quality VIS, and IDOL [88] specializes in online VIS.

F. Multi-object Tracking and Segmentation

Multi-object tracking (MOT) is a crucial task in autonomous driving, encompassing both object detection and tracking within video sequences. Numerous datasets have been curated with a focus on driving scenarios, including KITTI tracking [89], MOTChallenge [30], UA-DETRAC [90], PathTrack [91], and PoseTrack [92]. However, none of these datasets offer segmentation masks for annotated objects, thus lacking pixel-level representations and intricate interactions seen in MOTS data. More advanced datasets, such as Cityscapes [93], ApolloScape [94], BDD100K [95], and KITTI MOTS dataset [19], do provide instance segmentation data for autonomous driving. Nevertheless, Cityscapes only supplies instance annotations for a small subset (i.e., 5,000 images), and ApolloScape does not offer temporal object descriptions over time. Consequently, neither dataset is suitable for the joint training of MOTS algorithms. In contrast, KITTI MOTS [19] stands as the first public dataset that addresses the scarcity of data for the MOTS task, albeit with a relatively modest number of training samples. To date, BDD100K boasts the largest scale of data from intensive sequential frames, which may be considered redundant for training purposes. In comparison to the aforementioned datasets, our DGL-MOTS dataset encompasses a wider range of diverse data with finely detailed annotations.

III. THE PROPOSED METHOD

A. Overview

Fig. 1 shows the proposed network for multi-objects tracking and segmentation. It follows the basic network architectures for object detection, instance segmentation, and object tracking, containing two parts: one for extracting feature maps and the other for different tasks sharing the extracted features as inputs. More specifically, the first part of our method consists of two components: a backbone feature extraction network N_{fm} to compute per frame feature maps and an optical flow network N_{fl} to estimate object movements across nearby frames. With the help of optical flow, feature maps from previous frames can be warped into the current frame, which is used to combine with the extracted feature maps at the current frame, to obtain an enhanced feature representation of the current frame that should be more robust to image blur, occlusion, etc. Accordingly, the second part of our method is constituted of three heads, each of which corresponds to a specific task, i.e., object detection, instance segmentation, and multi-object tracking. The detection head contains classification head N_{cl} and bounding-box regression head N_{bb} while the instance segmentation is achieved by the mask prediction head N_{mk} . The tracking head N_{tr} aims to identify the same objects that appeared in multiple frames.

To train such a network with different heads end-to-end, existing methods resort to multi-task learning that simultaneously optimizes losses related to individual tasks. However, those methods ignore the intrinsic correlations among these tasks, which could benefit each other if used properly. In detail, each loss is only designed and optimized specifically for one task and there is no interaction between different tasks and their corresponding losses. In other words, the performance

of the instance segmentation task will not affect that of the object detection task and vice versa. In this case, we argue that the designs of losses can be optimized. Therefore, we propose collaborative multi-task learning to enable information propagation among these individual tasks when optimizing the total learning objective containing all these tasks. For this purpose, we introduce associative connections among detection, segmentation, and tracking heads so that the network training could be aware of the interactions among these three tasks. The introduced associative connections are shown by the red arrows in Fig. 1. Besides enabling information propagation among different task heads, our network is more efficient compared to previous methods. Existing MOTS methods [19], [96], [97] that also use three task heads encounter the problem of computing redundant feature representations, as they rely on proposal features for instance segmentation and tracking. With the help of associative connections, our segmentation and tracking heads only take the detected bounding boxes or masks for input, thus avoiding computing features on unrelated proposals. In addition, due to the added associative connections, the mask head and tracking head can use more accurate features extracted on object bounding boxes instead of the region proposals. Therefore, higher-quality instance segmentation and tracking results can be expected.

In the following subsections, we describe in detail the proposed method, including feature extraction network, associative connections, training, and inference, as well as the network architecture. Table I lists the main symbols used in this paper for the neatness of description.

B. Network Architecture

Our network architectures have general and flexible design options. We craft state-of-the-art architectures into the proposed methods for different visual tasks. Particularly, the proposed method has three contributing modules: (i) the CNN backbone architecture employed for feature extraction over the input frame, (ii) the optical flow architecture used for motion estimation across frames, and (iii) the task heads for classification, location regression (bounding box), mask generation, and object tracking. With associative connections, the mask head and tracking head are applied to object-aware ROI instead of using ROI from proposals.

1) *Feature extraction:* The ResNet-101 [98] is used as our backbone feature extraction network N_{fm} in this paper. According to the practice of using ResNet-101 as the backbone in Faster R-CNN, the outputs of its final convolutional layer $C4$ are feedforwarded to the task heads.

2) *Motion estimation:* The simple version of FlowNet [99] is used as the flow network N_{fl} for motion estimation across video frames. It is pretrained on the synthetic Flying Chairs dataset [99]. According to [24], [28], we have the input frame half-sized and the output stride 4. Therefore, the output resolution of the generated flow field is 1/8 of the original frame size. Since the output of the feature extraction network has a stride of 16, we use bilinear interpolation to further down-sample the flow field and scale the field by half, to match the resolution of extracted feature maps. The down-sample process is non-learnable as the bilinear interpolation is

TABLE I
SUMMARIZATION OF THE MAIN NOTATIONS USED IN THIS PAPER.

Notation	
\mathbf{N}_{fm}	Feature extraction network
\mathbf{N}_{fl}	Optical flow network
\mathbf{N}_{cl}	Classification head
\mathbf{N}_{bb}	Bounding box head
\mathbf{N}_{mk}	Mask head
\mathbf{N}_{tr}	Tracking head
F_t	Feature maps at time t
F'_t	Enhanced feature maps at time t
I_t	Input frame at time t
$F_{t-i \rightarrow t}$	Warped feature maps from time $t-i$ to t
$\Delta D_{t-i \rightarrow t}$	Object movement between t and $t-i$
$\omega_{k \rightarrow t}$	Weight for fusing feature of the k th frame at the time t
b^i	The bounding box of the i^{th} object
b^i_{mk}	The bounding box computed from the predicted mask of the i^{th} object
b^i_{wb}	Weighted bounding box of the i^{th} object connecting the detection and segmentation heads to the tracking head
\mathcal{L}_{cls}	Loss for classification
\mathcal{L}_{box}	Loss for bounding box regression
\mathcal{L}_{mask}	Loss for mask generation
\mathcal{L}_{track}	Loss for tracking
\mathcal{L}_{total}	Total loss for training

a parameter-free layer in the network and is also differentiated during training.

3) *Task heads*: For the task heads, we follow architectures presented in Faster R-CNN [31], Mask RCNN [5], and TrackR-CNN [19] for different tasks. For detection, we craft Faster R-CNN, a two-stage detector for object classification and bounding box regression. For mask prediction, we follow Mask RCNN by adding a fully convolutional mask prediction branch. For the tracking task, we include an association layer [19] to calculate the distance of 128-D identity vectors to track different objects across frames. Among these heads, there are associative connections to collect object-aware features and propagate them across these tasks. Introducing associative connections among other task heads to facilitate information flow interactively among several tasks is also the main contribution of this work. Compared to previous proposal-based features, this work uses object-aware features by leveraging associative connections, which shows better performance in predicting masks and tracking the same identities.

C. Feature Extraction Guided by Optical Flow

Given an input frame I_t at time t , the feature extraction process can be expressed as $F_t = \mathbf{N}_{fm}(I_t)$. Note that F_t are only intermediate feature maps, which will be enhanced with features from previous frames based on the estimated object movements by \mathbf{N}_{fl} . It is the enhanced feature maps being passed into the following heads regarding different tasks. Although various backbone networks can be used here for feature extraction, we use ResNet-101 [98] in this paper due to its popularity.

To enhance the feature representation of different objects on the input frame by leveraging on its previous frames, we exploit the temporal visual cues with the help of an optical flow network \mathbf{N}_{fl} . Thus, the enhanced feature representation is also called flow-guided features. The warped feature from time $t-i$ to t is denoted as:

$$F_{t-i \rightarrow t} = \mathcal{WP}(F_{t-i}, \Delta D_{t-i \rightarrow t}), \quad (1)$$

where \mathcal{WP} is the feature warping function to predict the feature maps at t based on the feature maps at $t-i$ (i.e., F_{t-i}) and the estimated object movement $\Delta D_{t-i \rightarrow t}$ between t and $t-i$ frames. The movement is predicted by the optical flow network, i.e., $\Delta D_{t-i \rightarrow t} = \mathbf{N}_{fl}(I_{t-i}, I_t)$. By modeling the feature map movements from nearby frames, we hope to improve the extracted features that may have been originally compromised by motion blur, defocus, or occlusion that often happened in the video domain.

Since in real scenarios processing a specific video frame can only rely on its previous frames, given an input frame I_t , we obtain a set of warped feature maps for each of its previous frames to compute the enhanced feature maps of the current frame I_t . Specifically, with a predefined temporal range n , each feature map of the previous frames in this range is warped into frame t according to Eq. (1), resulting in a set of predicted feature maps $\{F_{t-i \rightarrow t} | i \in [1, n]\}$. Note that we warp previous features to every current frame for feature fusion instead of keyframes or using dense aggregation [24], [28]. Due to the high efficiency of the used flow network (i.e., FlowNet [99]), this enables our method to extract strong features but still with affordable computational cost. Runtime analysis is provided in Section III-E2.

Given the set of warped feature maps $\{F_{t-i \rightarrow t} | i \in [1, n]\}$, the fused feature maps F'_t at the frame I_t is then computed by the weighted average of these warped features and F_t ,

$$F'_t = \sum_{k \in [t-n, t-1]} (\omega_{k \rightarrow t} \cdot F'_{k \rightarrow t}) + F_t, \quad (2)$$

where the weight $\omega_{k \rightarrow t}$ is adaptively computed based on the similarity between $F_{k \rightarrow t}$ and F_t . Among many choices for defining the similarity between these feature maps, we follow the implementation in [100] to use a shallow fully convolutional network to output embedding vectors for similarity

computation. That is, $\omega_{k \rightarrow t}$ is computed as,

$$\omega_{k \rightarrow t} = \exp\left(\frac{F_{k \rightarrow t}^e \cdot F_t^e}{|F_{k \rightarrow t}^e||F_t^e|}\right), \quad (3)$$

where F^e is the embedding vector of F outputted by the shallow fully convolutional network. Note that the dot products in Eq. (2) and Eq. (3) is element-wise multiplication and all the obtained weights are normalized so that $\sum_{k \in [t-1, t-n]} \omega_{k \rightarrow t} = 1$.

Compared to the existing architectures like TrackR-CNN [19] which naively uses heavy 3D convolutions for feature extraction and fusion, our method models the spatial movements of objects in successive video frames in a more reasonable fashion to enhance the extracted features, could improve the accuracy for upper-stream tasks. With the help of a lightweight optical flow network (i.e., FlowNet [99]) to predict the object movements, our method for feature extraction has much fewer parameters than 3D convolutions. Therefore, our method keeps a faster runtime than TrackR-CNN and its variants. In addition, by tuning the temporal range n , we can actively control the tradeoff between inferring speed and accuracy.

D. Associative Connections Across Tasks

After the above-mentioned flow-guided feature fusion, the obtained feature representation could be significantly enhanced. These enhanced feature maps are then fed into the individual upper-stream task heads. These tasks include object detection, instance segmentation, and object tracking. Although these tasks are intrinsically related, existing methods (e.g., TrackR-CNN [19] and CAMOT [101]) simply add different learning objectives together. This paper explores the relations among these tasks for improving multiple object tracking and segmentation accuracy. Under this motivation, we propose collaborative multi-task learning by adding associative connections among different tasks to facilitate information flow through different heads. Different from previous methods where different task heads are independent, the three heads in our method jointly worked together. To be concrete, as shown in Fig. 1, there are three associative connections across the detection, segmentation, and tracking heads. The first one is the associative connection between the output of bounding box regression in the detection head and the input of the mask prediction head. The second one is a connection to link the outputs of detection and mask heads. The third one is the associative connection between the combined outputs of detection and mask heads and the input of the tracking head. With this implementation, we achieve much lower computational complexity per instance than the mask head in TrackR-CNN, where the ROI-based operations are repeatedly performed for final dense predictions. For comparison, the associative connections facilitate information sharing across different task heads to keep a low computational cost. Meanwhile, owing to the interleaved information flow among all three heads, we can jointly optimize their objectives as a whole, while previous works optimize each task's objective independently since all the task heads are independent in their methods.

1) Connection from Detection to Segmentation: Due to the good performance of Mask R-CNN [5] for instance segmentation, we adopt its head architecture for object detection and segmentation. The detection and instance segmentation are based on region proposals, and our network training procedure contains two stages Mask R-CNN and Faster R-CNN [31]. In the first stage, a Region Proposal Network (RPN) takes the extracted features from the backbone and outputs a set of region proposals for object detection. In the second stage, for the detection head, an ROI pooling layer is used to generate region features for each proposal, which are then used to predict the object class and the related bounding box by the classification N_{cl} and regression N_{bb} heads respectively. Then, for the mask head, an ROI pooling layer is used to generate object-aware features based on the bounding boxes outputted from the detection head. These object-aware features are used by the following layers to conduct instance segmentation for producing object masks.

Adding a connection from the output of the detection head to the ROI pooling layer of the mask head (i.e., the red arrow between the bounding box and mask heads in Fig. 1) can explore more accurate features focused on objects compared to using the region proposals that have been widely used in the literature [19], [22]. This is because the output of the detection head is the object bounding boxes that are finer than the region proposals regarding the object locations. Therefore, our mask head can accurately fire on the pixels of instances. In other words, by adding an associative connection between the output of the detection head and the input of the mask head, we achieve a better behavior for mask generation since the used features could be more focused on the object itself.

The learning objective of our detection head is identical to that of Faster R-CNN,

$$\mathcal{L}_{cls} + \mathcal{L}_{box} = \frac{1}{D} \sum_{d \in D} \mathcal{L}'(p_d, p_d^*) + \frac{1}{D} \sum_{d \in D} \mathcal{L}''(t_d, t_d^*), \quad (4)$$

where D is the total number of detections in a video sequence, p_d is the predicted probability of an object, p_d^* is the corresponding ground-truth label, t_d is the coordinates of the predicted bounding box, t_d^* is the ground-truth for the corresponding bounding box.

Similarly, the instance segmentation head is trained using the following loss function,

$$\mathcal{L}_{mask} = \frac{1}{D} \sum_{d \in D} \mathcal{L}'''(m_d, m_d^*), \quad (5)$$

where m_d and m_d^* are predicted mask and the ground-truth mask.

Note that the novelty of our method lies in introducing the associative connections to interact among different tasks, not the individual learning objective for each task.

2) Connection from Detection and Segmentation to Tracking: The tracking head aims to establish correspondences of the same identities across successive frames. Existing works, such as TrackR-CNN, extensively extract ROI features for all region proposals, which are then used to compute identity vectors. Those identity vectors from positive proposals are used to link the same identities across frames based on their

similarities. Such a tracking head computes many redundant proposal features and identity vectors. In addition, extracting ROI features from region proposals may not be accurate enough to facilitate good tracking results. To address these issues, we propose to extract ROI features from the detection and instance segmentation results. This not only avoids the redundant feature computation but also provides a more accurate description of the tracked objects. Such an idea inspires the associative connection from the outputs of detection and segmentation heads to the input of the tracking head.

Usually, the bounding boxes generated by the detection head are larger than the ground truth boxes, while the bounding boxes computed from the masks produced by the mask head are often tighter than the ground truth ones due to the pixel-level prediction of the mask head. To effectively combine two kinds of bounding boxes for a better ROI feature extraction in the tracking head, we propose an adaptive fusion strategy to obtain a weighted bounding box as,

$$b_{wb}^i = \alpha_1 \cdot b^i + \alpha_2 \cdot b_{mk}^i, \quad (6)$$

where b^i is the bounding box outputted by the detection head and b_{mk}^i is the bounding box drawn from the predicted mask for the i^{th} object, α_1, α_2 are adaptive weights fulfilling $\alpha_1 + \alpha_2 = 1$. The parameters α_1 and α_2 are hyperparameters but they do not vary for each scenario right now. We will investigate how to adjust them for each video sequence in future work.

In this way, the weighted bounding boxes b_{wb} will be smaller than the bounding boxes b generated by the detection head, while larger than the mask-based bounding boxes b_{mk} . If the object is on a large scale, we expect the weighted bounding box to be tighter for tracking. However, if the object is small, the mask-based bounding box may not be accurate enough so we expect the weighted bounding box to approach the detected bounding box to include more information. Under this consideration, α_1 is set proportional to the reciprocal of the scale of the object, to be specific, the area of bounding box b^i . Therefore, we have a small α_1 (α_2 is large) and b_{mk}^i has a larger impact on the weighted bounding box b_{wb}^i when the object is large. On the contrary, b^i will have a larger impact on b_{wb}^i when the object is small.

Owing to the good properties of these adaptive bounding boxes, we add an associative connection to link these boxes into the tracking head for ROI feature extraction. In other words, we use b_{wb} to pool feature maps to be more object-aware. Following the idea of [19], [63] that use the embedding vectors to re-identify persons, our tracking head uses one fully connected layer to map ROI features into identity vectors v , each of which indicates a unique identified instance. These identity vectors v are used to link all detections across frames. Since our ROI features are extracted from object-aware feature maps and could be more discriminative for tracking multiple objects, our tracking head can predict more distinguishable identity vectors.

In the training stage, we minimize the distances of all vectors belonging to the same object while maximizing the distances of vectors belonging to different objects. To this end, the tracking head is trained with the batch hard triplet ranking

loss. For each detected object, we sample its hard positive detections and negative detections for network training. Let us denote D as all the detections in a video. At frame, t , each detection $d \in D$ is associated with a tracking vector v_d and a ground truth track ID which is used to determine its overlap with the ground truth over frames. Thus, for a video sequence including T frames, the tracking loss is computed by,

$$\begin{aligned} \mathcal{L}_{track} = & \frac{1}{D} \sum_{d \in D} \max \left(m + \max_{\substack{n \in D \\ id_d \neq id_n}} \mathcal{S}(v_d, v_n) \right. \\ & \left. - \min_{\substack{p \in D \\ id_d = id_p}} \mathcal{S}(v_d, v_p), 0 \right), \end{aligned} \quad (7)$$

where the subscript p and n indicate the positive and negative detections respectively and \mathcal{S} represents the similarity between the input vectors. By default, we use cosine similarity.

E. Inference and Training

In this section, we first elaborate on the inference algorithm, then analyze the runtime complexity, and finally discuss the training objective of the proposed method.

1) Inference algorithm: We summarize the inference of our proposed method in Algorithm 1. The algorithm of our proposed method can be divided into two major stages. In the first stage, given a video frame I_t , the feature extraction network \mathbf{N}_{fm} takes it as input and produces a set of intermediate feature maps F_t (line 2 in Algorithm 1). Based on the light-weight optical flow network \mathbf{N}_{fl} , we warp the temporal feature maps in previous frames to the current frame (line 3 to 4) according to Eq. (1). In the meantime, we calculate the similarities of the warped feature maps to that of the current frame. These similarities are used as adaptive weights (line 5) to perform flow-guided feature fusion over the temporally aligned features, to enhance the feature representation for the current frame. In this way, the enhanced feature maps F'_t (lines 6 to 7) can be obtained for different head tasks. In the second stage, the enhanced feature maps F'_t are fed into the classification and bounding box heads to predict object class $\{c_t\}$ and bounding box $\{b_t\}$ respectively (line 8). Using the predicted bounding box from $\{b_t\}$, we further obtain the object-aware feature maps $F'_{(bb)t}$ which are more accurate than the proposal-based features that are widely used in Mask RCNN [5] and its variants [19], [22]. Next, the feature maps $F'_{(bb)t}$ are fed into the mask head \mathbf{N}_{mk} (line 11). Since the mask head \mathbf{N}_{mk} takes the object-aware feature maps as inputs, it tends to generate masks (denoted as $\{m_t\}$) with higher quality compared to those masks predicted from proposal based features. We draw a close rectangular over each instance in $\{m_t\}$ to obtain a mask-based bounding box $\{b_{(mk)t}\}$. Considering the predictions from both $\{b_t\}$ and $\{b_{(mk)t}\}$ together, we compute adaptive weights to get a weighted bounding box $\{b_{(wk)t}\}$ for each detected object (line 12) which could be more accurate than the original bounding boxes $\{b_t\}$ produced by the bounding box head. Intuitively, more accurate bounding boxes can benefit following object tracking heads as they only focus on the major component of objects, reducing the influence of background. The final video recognition result R_t consists of a set of object classes $\{c_t\}$,

Algorithm 1 Online inference of the proposed method per frame

```

1: input: frame  $\{I_t\}$                                  $\triangleleft$  Video frame at time  $t$ 
2:  $F_t = \mathbf{N}_{fm}(I_t)$                            $\triangleleft$  Produce feature maps
3: for  $j = t - n$  to  $t - 1$  do
4:    $F_{j \rightarrow t} = \mathcal{WP}\left(F_j, \mathbf{N}_{fl}(I_j, I_t)\right)$      $\triangleleft$  Feature warping to time  $t$ 
5:    $\omega_{j \rightarrow t} = \exp\left(\frac{F_{j \rightarrow t}^e \cdot F_t^e}{|F_{j \rightarrow t}^e||F_t^e|}\right)$      $\triangleleft$  Calculate adaptive weights
6: end for
7:  $F'_t = \sum_{j=t-n}^{t-1} (\omega_{j \rightarrow t} \cdot F_{j \rightarrow t}) + F_t$      $\triangleleft$  Flow-guided feature fusion
8:  $\{c_t\} = \mathbf{N}_{cl}(F'_t)$                                  $\triangleleft$  Object detection results
9:  $\{b_t\} = \mathbf{N}_{bb}(F'_t)$                                  $\triangleleft$  bounding-box features maps
10:  $\{m_t\} = \mathbf{N}_{mk}(F'_{(bb)t})$                           $\triangleleft$  Produce mask results
11:  $b_{(mk)t} \leftarrow \{m_t\}$                                  $\triangleleft$  Mask-based bounding box
12:  $\{b_{(wb)t}\} = \sum_{i \in \mathbb{R}^n} \alpha_1 \cdot b_i^i + \alpha_2 \cdot b_{(mk)t}^i$      $\triangleleft$  Weighed bounding box
13:  $F'_{(tb)t} = \{b_{(wb)t}\} \rightarrow F'_t$                  $\triangleleft$  Object-aware features maps
14:  $\{t_t\} = \mathbf{N}_{tr}(F'_{(tb)t})$                           $\triangleleft$  Produce tracking results
15:  $R_t = (\{b_t\}, \{c_t\}, \{m_t\}, \{t_t\})$                    $\triangleleft$  Final recognition at  $t$ 
15: output: Video recognition result  $R_t$ .

```

bounding-box locations $\{b_t\}$, object masks $\{m_t\}$, and tracked identity labels $\{t_t\}$.

2) *Time complexity*: Based on Algorithm 1, we give an analysis of the time complexity of the proposed method. Besides the backbone feature extraction network \mathbf{N}_{fm} , there are five modules: (1) The optical flow network \mathbf{N}_{fl} which includes the bilinear warping and feature embedding functions; (2) The classification head \mathbf{N}_{cl} ; (3) The bounding-box regression head \mathbf{N}_{bb} ; (4) The mask head \mathbf{N}_{mk} ; (5) The tracking head \mathbf{N}_{tr} . Since our associative connections are parameters-free and only work as information flow which is fast, we exclude it in our analysis. Given a temporal range of n in flow-guided feature fusion, the optical flow network \mathbf{N}_{fl} loops n times per frame for feature warping. Accordingly, the runtime complexity for the proposed method is,

$$\mathcal{O}_{ours} = \mathcal{O}(\mathbf{N}_{fm}) + n \cdot \mathcal{O}(\mathbf{N}_{fl}) + \mathcal{O}(\mathbf{N}_{cl}) + \mathcal{O}(\mathbf{N}_{bb}) + \mathcal{O}(\mathbf{N}_{mk}) + \mathcal{O}(\mathbf{N}_{tr}) \quad (8)$$

We compare the time complexity of our method to its predecessor, TrackR-CNN [19] as both of them have similar architecture and TrackR-CNN is the state-of-the-art multi-object tracking and segmentation. With the same symbols, TrackR-CNN has the following runtime complexity,

$$\mathcal{O}_{tr} = \mathcal{O}(\mathbf{N}_{fm}) + m \cdot \mathcal{O}(\mathbf{N}_{3D}) + \mathcal{O}(\mathbf{N}_{cl}) + \mathcal{O}(\mathbf{N}_{bb}) + \mathcal{O}(\mathbf{N}_{mk}) + \mathcal{O}(\mathbf{N}_{tr}), \quad (9)$$

where $\mathcal{O}(\mathbf{N}_{3D})$ is the 3D convolutions and m is the feature fusion length used in TrackR-CNN for per frame feature extraction.

Typically, the backbone feature extraction network has a heavier architecture than the task heads because it contains many more layers for feature abstraction. Therefore, it is reasonable to assume $\mathcal{O}(\mathbf{N}_{cl}) \ll \mathcal{O}(\mathbf{N}_{fm})$, $\mathcal{O}(\mathbf{N}_{bb}) \ll \mathcal{O}(\mathbf{N}_{fm})$, $\mathcal{O}(\mathbf{N}_{mk}) \ll \mathcal{O}(\mathbf{N}_{fm})$, $\mathcal{O}(\mathbf{N}_{tr}) \ll \mathcal{O}(\mathbf{N}_{fm})$, and $\mathcal{O}(\mathbf{N}_{mk}) \ll \mathcal{O}(\mathbf{N}_{fm})$. Given these, the ratio of runtime complexity of the proposed method to TrackR-CNN [19] can be computed as:

$$C = \frac{\mathcal{O}_{ours}}{\mathcal{O}_{tr}} \approx \frac{\mathcal{O}(\mathbf{N}_{fm}) + n \cdot \mathcal{O}(\mathbf{N}_{fl})}{\mathcal{O}(\mathbf{N}_{fm}) + m \cdot \mathcal{O}(\mathbf{N}_{3D})}. \quad (10)$$

Compared to \mathbf{N}_{3D} used in TrackR-CNN, we use \mathbf{N}_{fl} which is more efficient while still being effective for object movement modeling. Assuming $n = m$, C in Eq. (10) is less than 1. Therefore, our method is faster than TrackR-CNN. Experimental results in Table II also demonstrate that our method is two times faster than TrackR-CNN.

3) *Training objective*: Since our method simultaneously considers object detection, segmentation, and tracking, its training objective contains multiple losses accordingly,

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask} + \mathcal{L}_{track}, \quad (11)$$

where \mathcal{L}_{cls} , \mathcal{L}_{box} , \mathcal{L}_{mask} , and \mathcal{L}_{track} denotes the loss for classification, bounding box regression, mask segmentation, and object tracking respectively. Following the implementation from Mask R-CNN [5], we modify the classification loss \mathcal{L}_{cls} and bounding-box loss \mathcal{L}_{box} based those from Faster R-CNN [31]. In detail, we add extra losses before and after the associative connections across tasks for full supervision. Therefore, both the predictions before and after the associative connections are optimized. As we use a set of $m \times m$ for mask generation, the mask head has a cm^2 -dimensional output for each ROI, which encodes c binary masks. To achieve this goal, we use a per-pixel sigmoid, and \mathcal{L}_{mask} is defined as the average binary cross-entropy loss. For an ROI associated with ground-truth class k , we only define \mathcal{L}_{mask} on the k^{th} mask while ignoring other mask outputs that do not contribute to the loss. The details of \mathcal{L}_{track} are discussed in Section III-D2.

As the entire framework of the proposed method is multiple-stage, we rely on predictions from the classification head for the label of object detection, instance segmentation (mask), and tracking. Using the classification result for each ROI, we allow each task head to generate its results. Namely, the bounding-box head, the mask head, and the tracking head only need to focus on their specific task based on ROI without competition among classes. With the associative connections, critical ROI information can forward pass from the bounding box head to the mask head and eventually to the tracking head respectively. In the same vein, the ground truth for instance segmentation and object tracking can be backpropagated to each task head in the training procedure.

IV. EXPERIMENTS

This section first describes the datasets and evaluation metrics that we use to assess our method. The implementation details with training parameter settings are then elaborated. Next, We supply ablation studies to investigate the improvements of our method from a strong baseline and its variants. We also compare our method with the state-of-the-art methods on two benchmarks. Finally, we conclude with the accuracy and runtime evaluation.

A. Datasets and Evaluation Metrics

KITTI MOTS [19] and MOTS Challenge [30] are used to evaluate the effectiveness of the proposed method. KITTI MOTS has 21 video sequences of 8,008 frames, while MOTSChallenge has four video sequences of 2,862 frames. For KITTI MOTS, it includes 11,420 pedestrian and 26,899 car instances. There are 26,894 pedestrian instances in the MOTS Challenge.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE KITTI MOTS DATASET.

Method	Detect + Segment	Speed (s)	FPS	Cars / Pedestrians			
				sMOTSA↑	MOTSA↑	MOTSP↑	IDS↓
CAMOT [101]	TRCNN	0.76	1.32	67.4 / 39.5	78.6 / 57.6	86.5 / 73.1	220 / 131
CIWT [102]	TRCNN	0.28	3.57	68.1 / 42.9	79.4 / 61.0	86.7 / 75.7	106 / 42
ReMOTS [21]	TRCNN + BB2SegNet	3.33 / -	0.30	70.4 / -	84.4 / -	- / -	231 / -
TrackR-CNN [19]	TRCNN	0.50	2.00	76.2 / 46.8	87.8 / 65.1	87.2 / 75.7	93 / 78
BePix [103]	RRC [104] + TRCNN	0.36	2.77	76.9 / -	89.7 / -	86.5 / -	88 / -
Stem-Seg [105]	TRCNN	0.32	3.12	72.7 / 50.4	83.8 / 66.1	87.2 / 77.7	76 / 14
Ours	TRCNN	0.27	3.70	76.7 / 47.9	88.2 / 65.3	88.5 / 76.1	62 / 32

Following TrackR-CNN [19], we use **MOTSA** to measure the accuracy for multi-object tracking and segmentation, **MOTSP** to measure the precision of mask-based multi-object tracking and segmentation results, and **sMOTSA** to evaluate the soft multi-object tracking and segmentation accuracy. They are defined as follows,

$$\begin{aligned} MOTSA &= \frac{|TP| - |FP| - |IDS|}{|M|}, \\ MOTSP &= \frac{|\widetilde{TP}|}{|TP|}, \\ sMOTSA &= \frac{\widetilde{TP} - |FP| - |IDS|}{|M|}, \end{aligned} \quad (12)$$

where TP , \widetilde{TP} , FP , and IDS are true positive, soft true positive, false positive, and tracking ID switch score respectively. We refer readers to [19] for more details about each notation. These metrics collectively measure the performance of a multi-object tracking and segmentation system by considering three tasks (i.e., object detection, instance segmentation, and multi-object tracking) together.

B. Implementation Details

The proposed method is implemented on a workstation with one NVIDIA RTX GPU. To have a fair comparison with existing methods, we follow the same experimental setup as in TrackR-CNN [19]. To be specific, the backbone feature extraction network ResNet-101 [98] is pretrained on COCO [106] and Mapillary [107] datasets. The optical flow network FlowNet is pretrained on the Flying Chair dataset [99]. During the training process, the weights of ResNet-101 and FlowNet are fixed, and the other weights related to different task heads (i.e., N_{bb} , N_{cl} , N_{mk} and N_{tr}) are updated by learning on the target dataset, i.e. KITTI MOTS or MOTS Challenge. We train our model for 40 epochs with a learning rate of 5×10^{-7} using the Adam [108] optimizer and mini-batch size of eight. The temporal range n in Eq. (2) is set to eight, i.e., eight adjacent frames are used for the flow-guided feature extraction.

For the KITTI MOTS benchmark, there are 21 videos in total and we use 12 of them for training and the remaining for testing, following the practice in TrackR-CNN. We randomly keep some training data for validation and choose the best model on the validation set for testing. For the MOTS Challenge benchmark, since there are only four video sequences in total, we use cross-validation to test the performance of different methods. To be specific, we leave one video sequence for evaluation and train the model on the three others on the

MOTS Challenge. This process is repeated four times and the average result is reported.

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

Results on KITTI MOTS. The results compared with the state-of-the-art methods on KITTI MOTS dataset are shown in Table II. For a fair comparison, we list the leading methods with detection and segmentation architectures based on TrackR-CNN, which is the same as ours. The best result for each metric is highlighted in the table. The results of our method in Table II is quite encouraging. For car recognition, our method achieves the best result on MOTSP and IDS. Especially, our method improves IDS over other methods by a significantly large margin, which leads to the second-best method by 26. sMOTSA and MOTSA of our method are on par with the best model, a.k.a., BePix [103]. Regarding speed, our method is faster than BePix. The fast inference speed of our method is due to its flow-guided feature extraction in which a lightweight flow network is used. Finally, for pedestrian recognition, our method is better than all the compared methods on these metrics.

To qualitatively demonstrate the improvements of our method, we visualize our results and compare them with TrackR-CNN. As shown in Fig. 2, our method is less prone to false predictions. In the left two columns of Fig. 2, TrackR-CNN in the top row produces a "car" prediction on the red minivan while our method in the bottom row does not. A minivan is not a labeled object in the KITTI MOTS dataset. For the right two columns of Fig. 2, TrackR-CNN produces a false detection on traffic signs, predicting them as "pedestrian". In these examples, our method performs accurate predictions for both scenarios. What is more, TrackR-CNN frequently encounters missing detection. For comparison, our method can constantly detect objects through video frames as demonstrated in Fig. 3. The improvement of our method over TrackR-CNN mainly stems from the reliable object-aware features, which are brought from our associative connections. With the help of these connections, our method produces better feature representations than TrackR-CNN before the task heads produce individual predictions, thus achieving better results.

We also observe that our method has improved performance on tracking and mask quality. We argue that the proposed associative connections benefit the mask and tracking predictions as for the two tasks, we use refined ROI features for the final prediction. In contrast, TrackR-CNN produces predictions using coarse features based on proposals.

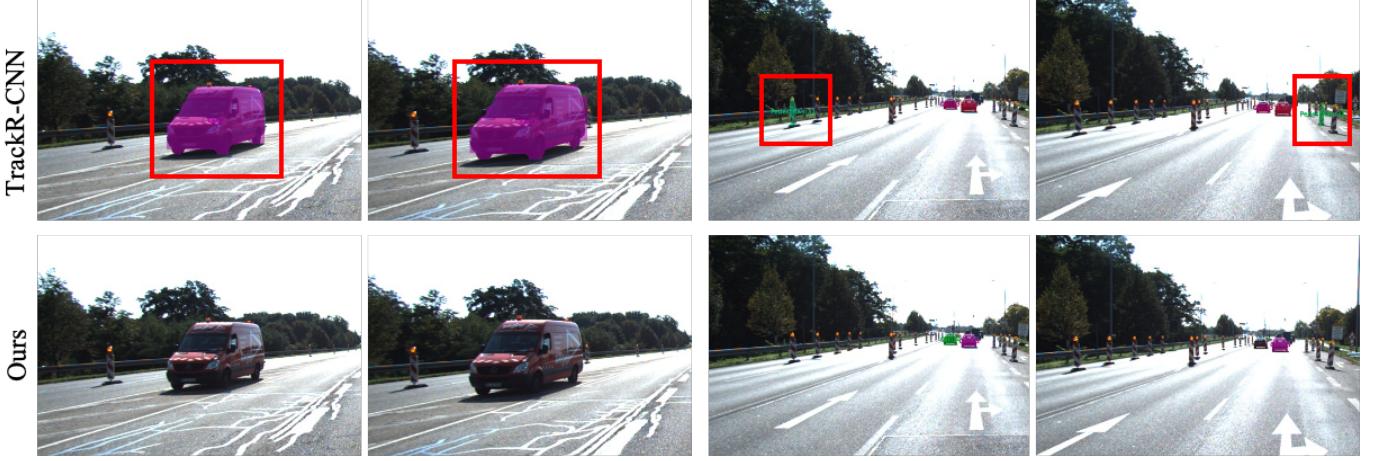


Fig. 2. Qualitative comparison between TrackR-CNN and our method regarding false prediction. The false predictions are marked by red bounding boxes. Minivan is not an object in the KITTI MOTS dataset. In the first two rows, TrackR-CNN has a false prediction on the minivan as a car class. In the same vein, TrackR-CNN has a false recognition of traffic signs and predicts them as "pedestrian". For both scenarios, our method works correctly without false predictions. The images are cropped for better visualization.

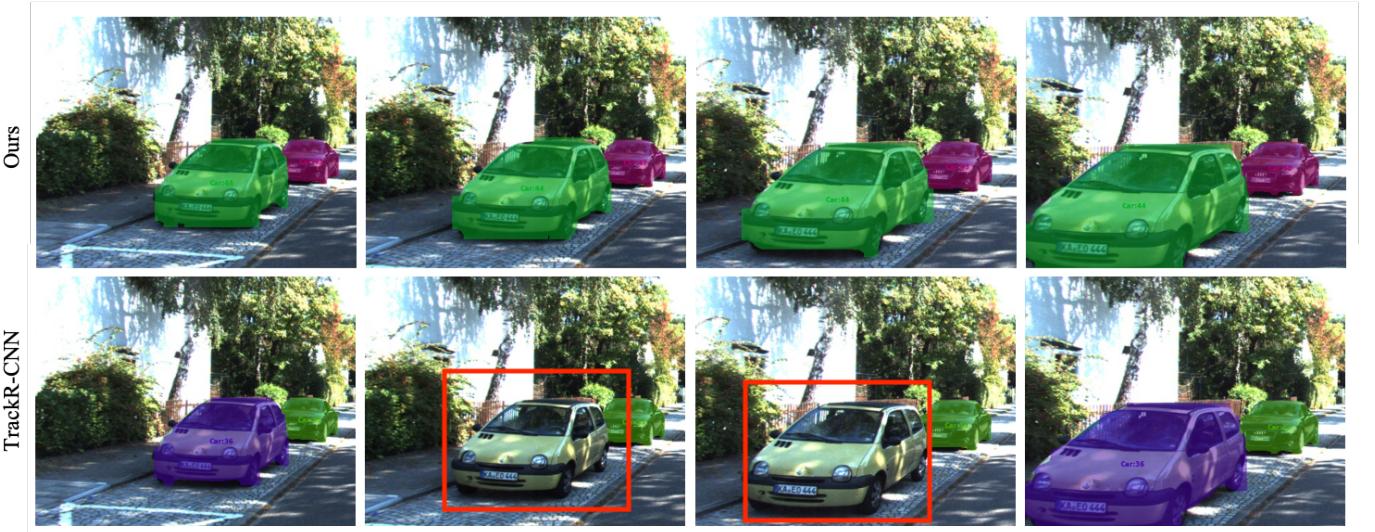


Fig. 3. Comparison of TrackR-CNN and our results on missed recognition. The missed recognition is marked in red bounding boxes. Our method can constantly perform recognition on video frames while TrackR-CNN encounters recognition lost for the middle two frames. The demonstrated examples are cropped for better visualization.

Results on MOTS Challenge. Table III reports the results on the MOTS Challenge dataset. Our method outperforms all the other methods for all metrics. The improvements over the second best method (TrackR-CNN), are 0.5 on sMOTSA, 0.2 on MOTSA, and 0.7 on MOTSP respectively. Besides TrackR-CNN, our method achieves significantly better results than other methods even if they use a domain-finetuned Mask R-CNN.

D. Ablation Study

To demonstrate the effectiveness of our design, we conduct ablation studies from TrackR-CNN since our method is proposed for the same purpose as TrackR-CNN and shares a similar architecture to it. Specifically, we first replace

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON MOTS CHALLENGE DATASET. +MG DENOTES MASK GENERATION WITH A DOMAIN FINETUNED MASK R-CNN.

Method	sMOTSA↑	MOTSA↑	MOTSP↑
MOTDT [109] + MG	47.8	61.1	80.0
MHT-DAM [110] + MG	48.0	62.7	79.8
jCC [111] + MG	48.3	63.0	79.9
FWT [112] + MG	48.3	64.0	79.7
TrackR-CNN [19]	52.7	66.9	80.2
Ours	53.2	67.1	80.9

the computationally expensive conv3d in TrackR-CNN with more efficient flow-guided feature extraction proposed in Section III-C, which leads to Method A. Then, we add an associative connection from the bounding box head to the

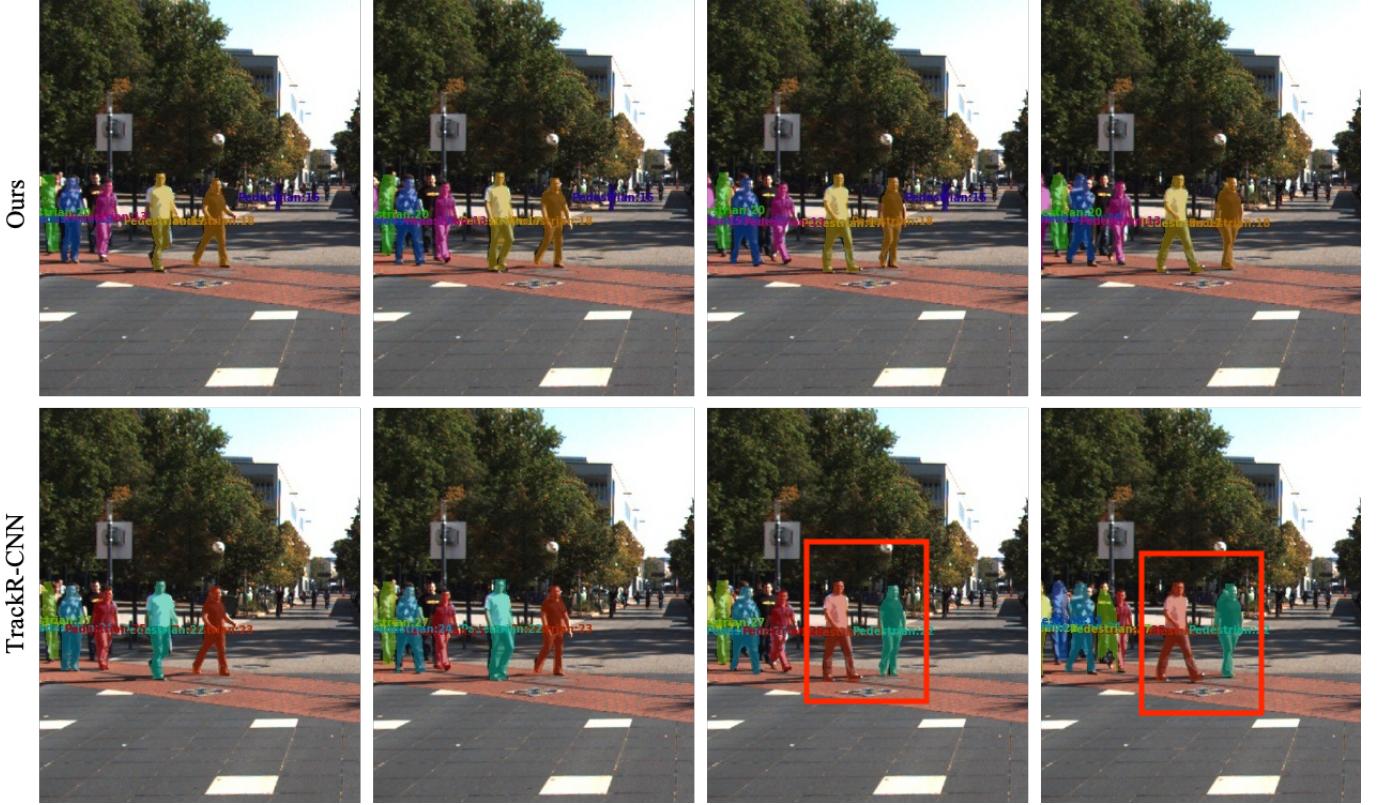


Fig. 4. Comparison of TrackR-CNN and our results on ID switch. The incorrect cases for tracking are marked in red bounding boxes. Our method can constantly perform accurate tracking of the same instances on video frames while TrackR-CNN frequently produces incorrect predictions. The demonstrated examples are cropped for better visualization.

mask head as described in Section III-D1 and obtain Method B. By further adding an associative connection from the bounding-box head to track, this leads to Method C in which bounding box features are used for both segmentation and tracking, rather than the proposal based features. A variant of Method C is to connect the mask-based bounding box to the tracking head, which we call Method D. We also test two kinds of strategies to fuse bounding boxes and mask-based bounding boxes, which are then input to the tracking head. Method E uses fixed weight $\alpha = 0.5$ to fuse these two kinds of bounding boxes, while Method F is the proposed one that fuses them with adaptive weights.

Table IV demonstrates the results of each method on the KITTI MOTS dataset. For reference, we also include the results of TrackR-CNN. By comparing A to TrackR-CNN, we can find that using flow-guided feature extraction leads to slightly worse accuracy, however, it leads to faster runtime as shown in Table II. Based on A, B obtains better results, proving our associative connection that using predicted bounding boxes is better than region proposals in pooling features for the mask head. C and D further improve B, demonstrating the effectiveness of introducing an associative connection to the tracking head. If these two kinds of bounding boxes together, better results are obtained. More importantly, the proposed adaptive box fusion strategy (Method F) achieves significant improvement over the hard fusion (Method E), especially for the tracking consistency. These results show that the proposed method effectively explores the properties of these two kinds for multi-object tracking and segmentation.

E. Fusion Length and Accuracy

We use eight frames as the default temporal range for feature fusion based on the object movements across frames. To evaluate the impact of this fusion length on the accuracy, we change the number of input frames for our method and test our method on the KITTI MOTS dataset. Table V reports the results concerning the number of input frames. For both cars and pedestrians, the recognition accuracy increases when the temporal range increases from 2 to 8. After that, the performance starts to degrade. Such results are reasonable, because only a few frames may not accumulate enough information to enhance the feature of the current frame while too many frames are also damageable as the long-term movements estimated by the optical flow network are unstable. In addition, it is also worth pointing out that a larger temporal range requires more time to process feature extraction, thus will result in a lower speed. Therefore, we need to pay more attention to the temporal range in practice to achieve a desirable recognition accuracy as well as the runtime speed.

V. CONCLUSION

In this work, we propose a novel algorithm capable of associatively detecting, segmenting, and tracking multiple objects for video analysis. By adding associative connections across detection, segmentation, and tracking heads in an end-to-end learnable CNN, our method enables information propagation through different tasks, which could benefit all the considered tasks. Therefore, our method achieves state-of-the-art performance on various metrics regarding object detection,

TABLE IV

ABALION STUDY ON THE KITTI MOTS DATASET. PLEASE SEE THE TEXTS FOR DETAILS ABOUT METHODS A TO E, AND F IS THE PROPOSED METHOD.

Method		TrackR-CNN	A	B	C	D	E	F
Cars	sMOTSA	76.2	75.6	76.0	76.1	76.5	76.3	76.5
	MOTSA	87.8	87.0	87.1	87.5	88.1	87.9	88.2
	MOTSP	87.2	86.9	87.1	87.4	88.4	88.3	88.5
	IDS	93	95	87	76	74	73	62
Pedestrians	sMOTSA	46.8	45.2	45.5	46.1	46.1	46.2	47.9
	MOTSA	65.1	63.5	63.8	64.3	64.3	64.4	65.3
	MOTSP	75.7	75.0	75.1	75.8	75.9	76.0	76.1
	IDS	78	76	43	39	38	36	32

TABLE V

INFLUENCE OF THE TEMPORAL RANGE USED IN THE FLOW-GUIDED FEATURE FUSION. THE RESULTS ARE ON THE KITTI MOTS DATASET.

temporal range n	2	4	8	12	16	
Cars	sMOTSA↑	75.2	75.7	76.7	75.4	75.3
	MOTSA↑	86.5	87.6	88.2	86.8	86.7
	MOTSP↑	87.3	87.4	88.5	87.8	87.2
	IDS↓	87	75	62	73	78
Pedestrians	sMOTSA↑	46.2	46.9	47.9	46.5	46.3
	MOTSA↑	63.5	64.3	65.3	63.7	63.6
	MOTSP↑	75.4	75.7	76.1	75.5	75.4
	IDS↓	39	36	32	38	35

instance segmentation, and multi-object tracking, according to our evaluation of two benchmarks. We also conducted extensive ablation studies to demonstrate the effectiveness of each associative connection.

REFERENCES

- [1] J. Liu, M. Gong, and H. He, “Deep associative neural network for associative memory based on unsupervised representation learning,” *Neural Networks*, vol. 113, pp. 41 – 53, 2019.
- [2] H. Luo, Y. Yang, B. Tong, F. Wu, and B. Fan, “Traffic sign recognition using a multi-task convolutional neural network,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1100–1111, 2018.
- [3] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, “Background prior-based salient object detection via deep reconstruction residual,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2015.
- [4] Z. Huang, X. Xu, H. He, J. Tan, and Z. Sun, “Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 4, pp. 730–741, 2019.
- [5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *ICCV*. Venice, Italy: IEEE, 2017, pp. 2961–2969.
- [6] D. Liu, Y. Cui, X. Guo, W. Ding, B. Yang, and Y. Chen, “Visual localization for autonomous driving: Mapping the accurate location in the city maze,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3170–3177.
- [7] L. Liu, Z. Dong, Y. Wang, and W. Shi, “Prophet: Realizing a predictable real-time perception pipeline for autonomous vehicles,” in *2022 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 2022, pp. 305–317.
- [8] L. Yan, Y. Cui, Y. Chen, and D. Liu, “Hierarchical attention fusion for geo-localization,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, virtual: IEEE, 2021, pp. 2220–2224.
- [9] D. Liu, Y. Cui, L. Yan, C. Mousas, B. Yang, and Y. Chen, “Densernet: Weakly supervised visual localization using multi-scale feature aggregation,” in *AAAI*. Virtual: AAAI, 2021, pp. 6101–6109.
- [10] D. Liu, Y. Cui, Z. Cao, and Y. Chen, “Indoor navigation for mobile agents: A multimodal vision fusion model,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE. virtual: IEEE, 2020, pp. 1–8.
- [11] Y. Wang, D. Liu, H. Jeon, Z. Chu, and E. T. Matson, “End-to-end learning approach for autonomous driving: A convolutional neural network model.” in *ICAART (2)*. Prague, Czechia: IEEE, 2019, pp. 833–839.
- [12] Z. Cao, Z. Chu, D. Liu, and Y. Chen, “A vector-based representation to enhance head pose estimation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. virtual: IEEE, 2021, pp. 1188–1197.
- [13] Z. Cao, D. Liu, Q. Wang, and Y. Chen, “Towards unbiased label distribution learning for facial pose estimation using anisotropic spherical gaussian,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, Springer. Tel Aviv, Israel: Springer, 2022, pp. 737–753.
- [14] L. Yan, Q. Wang, Y. Cui, F. Feng, X. Quan, X. Zhang, and D. Liu, “Gl-rg: Global-local representation granularity for video captioning,” in *International Joint Conference on Artificial Intelligence*, 2022.
- [15] L. Yan, S. Ma, Q. Wang, Y. Chen, X. Zhang, A. Savakis, and D. Liu, “Video captioning using global-local representation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6642–6656, 2022.
- [16] G. Bertasius and L. Torresani, “Classifying, segmenting, and tracking object instances in video with mask propagation,” in *CVPR*. Virtual: IEEE, 2020, pp. 9739–9748.
- [17] L. Yang, Y. Fan, and N. Xu, “Video instance segmentation,” in *ICCR*. Seoul, Korea: IEEE, 2019, pp. 5188–5197.
- [18] Y. Cui, Z. Cao, Y. Xie, X. Jiang, F. Tao, Y. V. Chen, L. Li, and D. Liu, “Dg-labeler and dgl-mots dataset: Boost the autonomous driving perception,” in *WACV*. Waikoloa, HI, USA: IEEE, 2022, pp. 58–67.
- [19] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, “Mots: Multi-object tracking and segmentation,” in *CVPR*. Long Beach, CA, USA: IEEE, 2019, pp. 7942–7951.
- [20] D. Liu, Y. Cui, W. Tan, and Y. Chen, “Sg-net: Spatial granularity network for one-stage video instance segmentation,” in *CVPR*. Virtual: IEEE, 2021, pp. 9816–9825.
- [21] F. Yang, X. Chang, C. Dang, Z. Zheng, S. Sakti, S. Nakamura, and Y. Wu, “Remots: Self-supervised refining multi-object tracking and segmentation,” *arXiv e-prints*, pp. arXiv–2007, 2020.
- [22] Z. Xu, W. Zhang, X. Tan, W. Yang, X. Su, Y. Yuan, H. Zhang, S. Wen, E. Ding, and L. Huang, “Pointtrack++ for effective online multi-object tracking and segmentation,” *arXiv preprint arXiv:2007.01549*, 2020.
- [23] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, “Fast online object tracking and segmentation: A unifying approach,” in *CVPR*. Long Beach, CA, USA: IEEE, 2019, pp. 1328–1338.
- [24] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, “Deep feature flow for video recognition,” in *CVPR*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 2349–2358.
- [25] Y. Cui, L. Yang, and D. Liu, “Dynamic proposals for efficient object detection,” *arXiv preprint arXiv:2207.05252*, 2022.
- [26] G. Wang, Z. Qin, S. Wang, H. Sun, Z. Dong, and D. Zhang, “Towards accessible shared autonomous electric mobility with dynamic deadlines,” *IEEE Transactions on Mobile Computing*, 2022.
- [27] Y. Cui, “Dfa: Dynamic feature aggregation for efficient video object detection,” *arXiv preprint arXiv:2210.00588*, 2022.
- [28] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, “Flow-guided feature aggregation for video object detection,” in *ICCV*. Venice, Italy: IEEE, 2017, pp. 408–417.
- [29] Y. Cui, L. Yan, Z. Cao, and D. Liu, “Tf-blender: Temporal feature blinder for video object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8138–8147.
- [30] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “Mot16: A benchmark for multi-object tracking,” *arXiv preprint arXiv:1603.00831*, 2016.
- [31] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NeurIPS*. Montreal, Canada: MIT Press, 2015, pp. 91–99.
- [32] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, “Object detection with deep learning: A review,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

- [33] Y. Cui, "Dynamic feature aggregation for efficient video object detection," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 944–960.
- [34] D. Liu, Y. Cui, Z. Cao, and Y. Chen, "A large-scale simulation dataset: Boost the detection accuracy for special weather conditions," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.
- [35] L. Yang, C. Yiming, and D. Liu, "Techniques for using dynamic proposals in object detection," Jul. 27 2023, uS Patent App. 17/581,423.
- [36] X. Zhang, Y. Cui, Y. Wang, M. Sun, and H. Hu, "An improved ae detection method of rail defect based on multi-level anc with vss-lms," *Mechanical Systems and Signal Processing*, vol. 99, pp. 420–433, 2018.
- [37] Z. Dong, Y. Lu, G. Tong, Y. Shu, S. Wang, and W. Shi, "Watchdog: Real-time vehicle tracking on geo-distributed edge nodes," *ACM Transactions on Internet of Things*, vol. 4, no. 1, pp. 1–23, 2023.
- [38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*. IEEE, 2014, pp. 580–587.
- [39] R. Girshick, "Fast r-cnn," in *ICCV*. Santiago, Chile: IEEE, 2015, pp. 1440–1448.
- [40] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *CVPR*. Long Beach, CA, USA: IEEE, 2019, pp. 821–830.
- [41] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NeurIPS*. Barcelona, Spain: MIT Press, 2016, pp. 379–387.
- [42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jun 2017.
- [43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [44] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Glasgow, UK: Springer, 2020, pp. 213–229.
- [46] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2021. [Online]. Available: <https://openreview.net/forum?id=gZ9hCDWe6ke>
- [47] Y. Cui, X. Liu, H. Liu, J. Zhang, A. Zare, and B. Fan, "Geometric attentional dynamic graph convolutional neural networks for point cloud analysis," *Neurocomputing*, vol. 432, pp. 300–310, 2021.
- [48] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *CVPR*. Virtual: IEEE, 2021, pp. 14454–14463.
- [49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.
- [50] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *CVPR*. Long Beach, CA, USA: IEEE, 2019, pp. 4974–4983.
- [51] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *CVPR*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 2359–2367.
- [52] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *ICCV*. Seoul, Korea: IEEE, 2019, pp. 9157–9166.
- [53] ——, "Yolact++: Better real-time instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [54] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," in *ECCV*. Glasgow, UK: Springer, 2020, pp. 1–18.
- [55] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in *ECCV*. Glasgow, UK: Springer, 2020, pp. 649–665.
- [56] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," in *NeurIPS*. virtual: MIT Press, 2020, pp. 17721–17732.
- [57] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *ICCV*. Virtual: IEEE, 2021, pp. 6910–6919.
- [58] X. Liu, D. Tao, M. Song, L. Zhang, J. Bu, and C. Chen, "Learning to track multiple targets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1060–1073, 2015.
- [59] X. Wang, B. Fan, S. Chang, Z. Wang, X. Liu, D. Tao, and T. S. Huang, "Greedy batch-based minimum-cost flows for tracking multiple objects," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4765–4776, 2017.
- [60] P. Chu and H. Ling, "Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," in *ICCV*. Long Beach, CA, USA: IEEE, 2019, pp. 6172–6181.
- [61] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [62] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear lstm," in *ECCV*. Springer, 2018, pp. 200–215.
- [63] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *CVPR*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 3539–3548.
- [64] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *CVPR*. Las Vegas, Nevada, USA: IEEE, 2016, pp. 817–825.
- [65] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.
- [66] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang *et al.*, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2017.
- [67] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9217–9225.
- [68] Y. Cui, L. Yang, and H. Yu, "Dq-det: Learning dynamic query combinations for transformer-based object detection and segmentation," *arXiv preprint arXiv:2307.12239*, 2023.
- [69] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7210–7218.
- [70] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *CVPR*. Virtual: IEEE, 2020, pp. 10337–10346.
- [71] S. Wang, Y. Zhou, J. Yan, and Z. Deng, "Fully motion-aware network for video object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 542–557.
- [72] Y. Cui, "Feature aggregated queries for transformer-based video object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6365–6376.
- [73] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *ICCV*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 686–695.
- [74] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *CVPR*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 2663–2672.
- [75] B. B. Elallid, S. E. Hamdani, N. Benamar, and N. Mrani, "Deep learning-based modeling of pedestrian perception and decision-making in refuge island for autonomous driving," in *Computational Intelligence in Recent Communication Networks*. Springer, 2022, pp. 135–146.
- [76] L. Yang, Y. Wang, X. Xiong, J. Yang, and A. K. Katsaggelos, "Efficient video object segmentation via network modulation," in *CVPR*. Salt Lake City, UT, USA: IEEE, 2018, pp. 6499–6507.
- [77] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," *arXiv preprint arXiv:2007.14772*, 2020.
- [78] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *ICCV*. Seoul, Korea: IEEE, 2019, pp. 9627–9636.
- [79] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *ECCV*. Glasgow, UK: Springer, 2020.
- [80] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2former for video instance segmentation," *arXiv preprint arXiv:2112.10764*, 2021.
- [81] M. Heo, S. Hwang, S. W. Oh, J.-Y. Lee, and S. J. Kim, "Vita: Video instance segmentation via object token association," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23109–23120, 2022.

- [82] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, “End-to-end video instance segmentation with transformers,” in *CVPR Virtual*: IEEE, 2021, pp. 8741–8750.
- [83] S. Yang, X. Wang, Y. Li, Y. Fang, J. Fang, W. Liu, X. Zhao, and Y. Shan, “Temporally efficient vision transformer for video instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2885–2895.
- [84] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, “Video instance segmentation using inter-frame communication transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 352–13 363, 2021.
- [85] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, “Seqformer: Sequential transformer for video instance segmentation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 553–569.
- [86] L. Ke, H. Ding, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu, “Video mask transfiner for high-quality video instance segmentation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 731–747.
- [87] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, “Mask transfiner for high-quality instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4412–4421.
- [88] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, “In defense of online models for video instance segmentation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 588–605.
- [89] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.
- [90] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, “Ua-detrac: A new benchmark and protocol for multi-object detection and tracking,” *arXiv preprint arXiv:1511.04136*, 2015.
- [91] S. Manen, M. Gygli, D. Dai, and L. Van Gool, “Pathtrack: Fast trajectory annotation with path supervision,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 290–299.
- [92] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, “Posetrack: A benchmark for human pose estimation and tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5167–5176.
- [93] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset,” in *CVPR Workshop on the Future of Datasets in Vision*, vol. 2, 2015.
- [94] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, “The apolloscape dataset for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.
- [95] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving dataset for heterogeneous multitask learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2636–2645.
- [96] J. Luiten, I. E. Zulfikar, and B. Leibe, “Unovost: Unsupervised offline video object segmentation and tracking,” in *WACV*. Snowmass Village, CO, USA: IEEE, 2020, pp. 2000–2009.
- [97] H. Lin, X. Qi, and J. Jia, “Agss-vos: Attention guided single-shot video object segmentation,” in *ICCV*. Seoul, Korea: IEEE, 2019, pp. 3949–3957.
- [98] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*. Honolulu, Hawaii, USA: IEEE, 2016, pp. 770–778.
- [99] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, “Flownet: Learning optical flow with convolutional networks,” in *ICCV*. Santiago, Chile: IEEE, 2015, pp. 2758–2766.
- [100] D. Liu, Y. Cui, Y. Chen, J. Zhang, and B. Fan, “Video object detection for autonomous driving: Motion-aid feature calibration,” *Neurocomputing*, pp. 1–11, 2020.
- [101] A. Osep, W. Mehner, P. Voigtlaender, and B. Leibe, “Track, then decide: Category-agnostic vision-based multi-object tracking,” in *ICRA*. Brisbane, Australia: IEEE, 2018, pp. 1–8.
- [102] A. Osep, W. Mehner, M. Mathias, and B. Leibe, “Combined image-and world-space tracking in traffic scenes,” in *ICRA*, IEEE. Marina Bay Sands, Singapore: IEEE, 2017, pp. 1988–1995.
- [103] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna, “Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking,” in *ICRA*. Brisbane, Australia: IEEE, 2018.
- [104] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, “Accurate single stage detector using recurrent rolling convolution,” in *CVPR*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 5420–5428.
- [105] A. Athar, S. Mahadevan, A. Osep, L. Leal-Taixé, and B. Leibe, “Stemseg: Spatio-temporal embeddings for instance segmentation in videos,” *arXiv preprint arXiv:2003.08429*, 2020.
- [106] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft coco: Common objects in context,” 2014.
- [107] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *ICCV*. Venice, Italy: IEEE, 2017.
- [108] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014.
- [109] L. Chen, H. Ai, Z. Zhuang, and C. Shang, “Real-time multiple people tracking with deeply learned candidate selection and person re-identification,” in *ICME*, 2018.
- [110] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, “Multiple hypothesis tracking revisited,” in *ICCV*. Santiago, Chile: IEEE, Dec. 2015.
- [111] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, “Motion segmentation and multiple object tracking by correlation co-clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 140–153, 2020.
- [112] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, “Fusion of head and full-body detectors for multi-object tracking,” in *CVPR workshops*. Salt Lake City, UT, USA: IEEE, 2018, pp. 1428–1437.