A Unified Model for Tracking and Image-Video Detection Has More Power

Peirong Liu¹ *  Rui Wang²  Pengchuan Zhang²  Omid Poursaeed²  Yipin Zhou²
Xuefei Cao²  Sreya Dutta Roy²  Ashish Shah²  Ser-Nam Lim²
¹UNC-Chapel Hill  ²Meta AI

Figure 1. TrIVD enables image/video object detection and multi-object tracking within a single model. With the proposed unified framework, we are uniquely able to conduct zero-shot multi-object tracking on objects (airplanes, pandas, etc.) that have not appeared in tracking datasets. (Different colors refer to object identities in tracking and different object categories in detection figures.)

Abstract

Object detection (OD) has been one of the most fundamental tasks in computer vision. Recent developments in deep learning have pushed the performance of image OD to new heights by learning-based, data-driven approaches. On the other hand, video OD remains less explored, mostly due to much more expensive data annotation needs. At the same time, multi-object tracking (MOT) which requires reasoning about track identities and spatio-temporal trajectories, shares similar spirits with video OD. However, most MOT datasets are class-specific (e.g., person-annotated only), which constrains a model’s flexibility to perform tracking on other objects. We propose TrIVD (Tracking and Image-Video Detection), the first framework that unifies image OD, video OD, and MOT within one end-to-end model. To handle the discrepancies and semantic overlaps across datasets, TrIVD formulates detection/tracking as grounding and reasons about object categories via visual-text alignments. The unified formulation enables cross-dataset, multi-task training, and thus equips TrIVD with the ability to leverage frame-level features, video-level spatio-temporal relations, as well as track identity associations. With such joint training, we can now extend the knowledge from OD data, that comes with much richer object category annotations, to MOT and achieve zero-shot tracking capability. Experiments demonstrate that TrIVD achieves state-of-the-art performances across all image/video OD and MOT tasks.

* Work done during an internship at Meta AI.
Correspondence to: peirong@cs.unc.edu.
1. Introduction

Object detection (OD) consists of a localization and classification stage, in which the former determines the location of a potential object and the latter predicts the detected object’s category. Traditional detectors address this problem indirectly, by defining surrogate regression and classification problems on a large number of predicted proposals [10, 82], anchors [62], or window centers [94, 117]. Their performance therefore largely depends on the post-processing steps, e.g., approaches to collapse near-duplicate predictions, design the anchor sets or assign the target boxes to anchors [111]. DETR-based methods [12, 41, 74, 120], as fully end-to-end object detectors, were proposed to eliminate the need for hand-crafted components via the relation modeling capability of vision transformers (ViT) [24]. Coupled with language encoders and contrastive learning [81], recent open-vocabulary detection models are further able to leverage information from the large amounts of image/object-text data to boost the model’s performance and further achieve zero-shot capabilities [33, 47, 58, 74, 108, 110, 115].

However, the above developments mainly focus on image OD, leaving video OD less-explored, largely because video OD models usually have many bespoke hand-crafted components, e.g., optical flow [103, 104, 119], which requires prior knowledge from additional flow data. Another challenge lies in applying advance modern architecture like ViT on the high-resolution, space-time video OD data due to the high computational cost incurred by self-attention’s quadratic complexity [96]. In [41], the authors use a ViT to extract frame-level features first, and then apply another ViT to leverage the temporal relations. However, this strategy still faces quadratically increasing self-attention computations w.r.t. input videos’ temporal lengths. We instead formulate the image and video inputs in a single framework via decomposed temporal-aware attention [Sec. 3.1], with only linear computation increase along the temporal axis.

Meanwhile, multi-object tracking (MOT) models the tracking identities and the spatio-temporal trajectories [71], and shares similar goals with OD in general on locating potential objects, and with video OD in particular on reasoning about spatio-temporal relations between adjacent frames. Recent advances in MOT approaches mainly pursue tracking by detection [16, 42, 50, 51, 56, 57], by regression [6, 9, 20, 65, 116], or by attention [71, 105, 109, 112]. Built upon the recent advances of ViTs, tracking-by-attention [71, 120] associates objects across frames via the self-attention mechanism intrinsically introduced by ViTs [24, 96], and naturally relates tracking with frame-level detection. Our model follows tracking-by-attention mechanism. By inheriting proposed objects from previous frames, we achieve detection and tracking association simultaneously (Sec. 3.3).

We present a unified framework, TrIVD (Tracking and Image-Video Detection), which incorporates the three (image/video OD, MOT) tasks in one end-to-end model. TrIVD could be trained on image/video OD and MOT datasets separately, or co-trained in a cross-dataset, multi-task fashion. We highlight our contributions as follows:

Bridging the gap between image OD and video OD. Existing OD models are specifically designed for either image or video OD task with insufficient flexibility in handling inputs containing both images and videos. TrIVD formulates image and video inputs uniformly, with an integrated temporal-aware attention module to efficiently leverage the spatio-temporal relations for video inputs (Sec. 3.1).

Connecting MOT with image OD and video OD. We formulate MOT in a tracking-by-attention fashion [71], where detection and tracking data association are performed jointly via self-attention without additional track matching procedures (Sec. 3.3). This formulation enables the multi-dataset, multi-task training of TrIVD, and equips TrIVD with zero-shot tracking ability to track objects that have not been seen during MOT training (Fig. 1, Fig. 3).

One multi-dataset classifier with region-text alignment. Class categories vary across different OD/MOT datasets, yet semantic overlaps may exist. We re-formulate the class prediction in OD/MOT via phrase grounding [58] such that TrIVD is given both image/video and a text prompt containing all the candidate categories to be detected/tracked (Sec. 3.2). By aligning visuals with their semantic meanings, we intrinsically resolve the class label discrepancies and semantic overlaps across datasets.

TrIVD achieves state-of-the-art results across image OD, video OD and MOT (Sec. 4.3). Trained in a multi-dataset, multi-task fashion, we show that our unified model, uniquely achieves zero-shot tracking performance, and is able to track objects without the need for training with their ground truth tracking identity annotations (Fig. 1, Fig. 3).

2. Related Work

Image Object Detection (Image OD) aims to detect objects with their associated categories [43]. With the advent of convolutional neural networks (CNNs), current leading object detectors are built upon CNNs [11, 40, 54, 89, 93] and can be generally classified into two main categories: anchor-based detectors (e.g., R-CNN [31], Fast(er) R-CNN [30, 82], Cascade R-CNN [10], etc.) and anchor-free detectors (e.g., CornerNet [55], ExtremeNet [118], etc.). The former can be further divided into two-stage and one-stage methods, while the latter falls into the class of keypoint-based and center-based methods [111].

Recently, Transformers [12, 24, 71, 92, 99, 120] have received great attention in computer vision. DETR-based methods [12, 120] build a fully end-to-end object detection model based on Transformers, and largely simplify the traditional detection pipeline [82]. By coupling with language encoders and contrastive learning [81], a new stream
of open-vocabulary detection works are able to take advantage of the large amounts of image/object-text grounding data and further boost the model’s performance and achieve zero-shot capabilities [33, 47, 58, 74, 108, 110, 115]. TrIVD builds upon deformable-DETR [120] and resorts to region-text alignment for a unified classifier, resulting in an end-to-end, unified model for image/video OD and MOT (Fig. 1).

**Video Object Detection (Video OD)** requires not only detecting objects in each frame as image object detection, but also linking the same objects across frames. One common solution [15, 17, 34–36, 39, 46, 61, 91, 106] is using feature aggregation to enhance per-frame features by aggregating the features of nearby frames with flow-based warping [25, 103, 104, 119]. Another line of attention-based approaches utilize self-attention [96] and non-local information [98] to capture long-range dependencies of temporal contexts [8, 17, 21, 23, 39, 45, 102]. Despite making great progress, most pipelines for video OD are sophisticated and include multiple postprocessing steps as well as hand-crafted components [1, 5, 37, 48]. TransVOD [41] applies vision transformers (ViT) to build an end-to-end video OD model, and handles the spatio-temporal relations by using a ViT to extract frame-level features, and an additional ViT for temporal aggregation, which results in a quadratic increase in self-attention computation along the temporal axis. In contrast, TrIVD uniformly formulates image and video inputs with the proposed temporal-aware attention mechanism to efficiently fuse features across video frames (Sec. 3.1).

**Multi-object Tracking (MOT)** models the spatio-temporal trajectories of tracking identities [71]. Recent works generally focus on three aspects, tracking by detection, by regression, or by attention. Tracking-by-detection tackles MOT by detecting objects frame-wise and then associating the object identities across adjacent frames [16, 42, 50, 51, 56, 57]. Tracking-by-regression applies a continuous regression following the positions of detected objects between frames [6, 9, 20, 65, 116]. Tracking-by-attention associates objects via the self-attention [24, 71, 105, 109, 112], and naturally relates frame-level tracking and detection. We follow tracking-by-attention approaches, and integrate detection and tracking in our unified framework (Sec. 3.3). Co-trained on image/video OD and MOT datasets, TrIVD not only achieves state-of-the-art performance across all the three tasks, but is able to track novel object categories without the need for supervised training on their tracking annotations (Fig. 3).

**Multi-dataset, Multi-modal and Multi-task Learning** Multi-modal learning architectures allow training separate encoders for different input modalities, such as image-text [14, 32, 49, 69, 72], video-audio [2, 3, 76–79] and video-optical flow [88]. Most multi-modal models assume the input modalities are in correspondence and available simultaneously, while TrIVD operates on multi-modal inputs but yet does not require simultaneous access to all modalities.

Multi-task learning [13] operates on the same input but output predictions for multiple tasks [26, 28, 52, 70, 75, 113], while our model is able to handle both image and video inputs, and conduct OD or MOT tasks simultaneously (Fig. 1).

### 3. TrIVD: Tracking & Image-Video Detection

In this section, we introduce TrIVD, the unified tracking and image-video object detection framework. In Sec. 3.1, we first describe our unified formulation for image and video inputs, and propose the temporal-aware attention mechanism, as an efficient spatio-temporal feature aggregation module for video inputs. In Sec. 3.2, we introduce our unified classifier via region-text alignment for cross-dataset co-training, to handle the discrepancy and semantic overlaps across object categories from different datasets. In Sec. 3.3, we detail and conclude TrIVD’s entire unified framework for tracking and image-video detection.

#### 3.1. Unified Image-Video Backbone with Temporal-Aware Deep Fusion

We first introduce our formulation for unifying image and video inputs when extracting features from the backbone, then propose the temporal-aware attention module for video inputs.

**Image/Video Inputs** We represent videos as a set of frames and reshape the temporal dimension ($T$) into the batch dimension ($B$) to obtain a tensor $X \in \mathbb{R}^{B' \times H \times W \times C}$ in which $B' = B \times T$ is the new batch size, $H \times W$ refers to the spatial dimensions, and $C$ is the channel dimension. Similarly, we represent images as $X \in \mathbb{R}^{B \times H \times W \times C}$.

**Backbone** While our unified framework can use any vision transformer architecture [24] to process the image and video inputs, we adopt the MViTv2 [60] architecture as the backbone, which hierarchically expands the feature complexity while reducing the spatial resolution via attention-pooling, giving its proven advantage given its better performance and efficiency over single-scale vision transformers for image and video tasks [27, 60].

With our unified input formulation, the backbone model maps the input 2D patches into a shared representation $\Phi$ for both images and videos, using a 2D linear layer followed by LayerNorm [4]. Same embedding layers are also applied to embed all input (image/video) patches to enable maximal parameter sharing across the two visual modalities. Note that since all inputs are treated as single-frame images, only relative positional encoding [87] on the spatial domain is needed for either images or videos.

Therefore, the frame-level multi-scale features extracted by MViTv2 [60] are a set of 3D features,
Sequential Attentions for Video Inputs (Sec. 3.1)

$$F = \{F_l\} \quad (\text{Eq. (2)})$$

Region-Text Alignment (Sec. 3.2)

Figure 2. Overview of TrIVD’s unified framework. With a unified backbone, TrIVD could take both images and videos as inputs, using the proposed temporal-aware attention for spatio-temporal feature fusion of video features (Sec. 3.1). Depending on specific tasks, TrIVD performs object detection or tracking (Sec. 3.3). For detection in the unified detection-tracking context, we initialize our transformer decoder with empty object queries (white boxes). In tracking tasks, we initialize the object queries of current frame in combination with those of previous frames.

The temporal dimension of videos. Specifically, given a set of multi-scale features $F$ (Eq. (2)), we decompose the overall attention function $\pi$ over the space-time domain, i.e., $W(F) = \pi(F) \cdot F$, is decomposed into two sequential attentions along spatial and temporal axes:

$$W(F_l) = \pi_l(F_l) \cdot F_l, \quad l = 1, ..., L, \quad (3)$$

where $\pi_l(\cdot), \pi_{l,W_l}(\cdot)$ are two attention functions applied on the temporal axis ($T$), and spatial axis ($H_l \times W_l$) respectively (Fig. 2). We follow [19] for its attention design on the spatial domain ($\pi_{H_l,W_l}(\cdot)$), and apply an additional temporal attention module by dynamically aggregating features across their temporal dimensions:

$$\pi_l(F_l) \cdot F_l = \frac{1}{T} \cdot \sigma \left( f \left( \frac{1}{H_l W_l} \sum_{H_l W_l} F_l \right) \right), \quad l = 1, ..., L. \quad (4)$$

where $f(\cdot)$ is a linear function approximated by a $1 \times 1$ convolutional layer and $\sigma$ is the hard-sigmoid function.

Passing through the two sequential attention modules across spatial and temporal dimensions, we efficiently aggregate the spatio-temporal features from video inputs, and also achieve unified feature representations for images and videos, whose extracted features could be further forwarded to any downstream task-specific model.

3.2. Unified Cross-dataset Classifier via Grounding

One major task of detection/tracking is the class prediction for each proposed bounding box indicating the detected object category. Typical detection/tracking models predict the object class using a linear activation following the

_____

1In fact, we first considered Omnivore [29]’s video formulation as our unified image-video representation, yet we find the proposed frame-level feature extraction followed by temporal-aware deep fusion works better.

**Temporal-aware Attention**

One main challenge of spatio-temporal feature aggregation is the trade-off between performance and computational cost. Dai et al. [19] propose dynamic head for image OD, which decomposes attention on individual feature channels and improves the model’s efficiency. We further extend this sequential attention idea to
extracted bounding box features, which is usually trained with the multi-class cross entropy loss or focal loss [63]. However, the above classification losses defined on logit-encoded class labels are not easily generalizable in the case of multi-dataset joint-training. Typically, annotated object class labels vary across different OD/MOT datasets, yet semantic overlaps may exist among them.

One workaround is joint training with dataset-specific classification layers [29], but this could potentially result in conflicts due to the non-exhaustive annotations across varying datasets. A more elegant solution is binary classification with sigmoid activation [115], where the judgement for every object category is independent from others. When the ground-truth categories are from other datasets, the related logits are simply masked out during gradient backpropagation. Yet this strategy still requires re-arranging the class labels each time a new dataset is added to the training.

Aiming for a more generalized and flexible approach balancing the mixed object categories and their semantic overlaps, we re-formulate the cross-dataset classification problem as phrase grounding [58, 110], i.e., instead of classifying within $C$ classes, the class prediction is now achieved by aligning each proposed region to words in a text prompt. Specifically, during co-training, for each sample, we concatenate all available object categories in its beloved dataset to form a text prompt. For instance, $VID$ [84] dataset has the label to classname correspondences,

$$C_{VID} = \{1: \text{airplane}, 2: \text{antelope}, ..., 30: \text{zebra}\},$$

then the text prompt associated with samples from $VID$ is

$$T_{VID} = \text{“airplane antelope ... zebra”},$$

where each object class is converted to a candidate phrase to be grounded/aligned, parsed by blank spaces. Therefore, unlike the classification setting in typical detection/tracking models, $TrIVD$ does not directly output a class label for the proposed region/object, but assesses the token positions (soft tokens) that align with the regions/objects (Fig. 2).

To this end, we replace the regular classification loss with 1) the soft token loss ($L_{\text{soft}}$) [47], to encourage the predicted token spans to be aligned with the objects’ semantic meanings, and 2) the contrastive alignment loss ($L_{\text{contrastive}}$) [47, 58], to increase the similarities between the visual representations of the proposed objects, and the representations of the matched words in the text prompt.

### 3.3. Unified Tracking and Image-Video Detection

As illustrated in Fig. 2, the proposed $TrIVD$ consists of two major components: 1) Modality-agnostic visual feature extraction in the backbone, where one could opt to conduct frame-level feature extraction for images, or add spatial-temporal fusion for video inputs with the introduced temporal-aware attention module (Sec. 3.1); 2) Task-specific detection or tracking as follows.

### DETRs

Our detector/tracker is built upon the end-to-end detection framework Deformable DETR [12, 120], as its self-attention mechanism could be simultaneously adopted for object detection as well as tracking data association (Sec. 3.3). Briefly speaking, with a transformer encoder-decoder structure [96], Deformable DETR is initialized with a certain number ($N_{\text{box}}$) of object bounding boxes (i.e., object queries), to detect potentially existing objects in the boxes. Forwarding through the cross-attention modules in the transformer’s decoder, the model outputs the final predictions on the box coordinates along with their associated class label and confidence score. Deformable DETR is trained with the Hungarian matching loss, where a bipartite matching is computed between the $N_{\text{box}}$ predicted object queries and the ground-truth objects. The matched objects are encouraged to align with the ground-truth, while the unmatched ones are treated as background. Cross-entropy loss is used for classification supervision, $L_1$ loss and Generalized IoU are used for the bounding box supervision.

### Detection-Tracking Bipartite Matching

Since $TrIVD$ does not directly predict class labels, but aligns the token positions in the text prompt with the proposed object (Sec. 3.2), the bipartite matching between the ground truth and proposed objects do not rely on class labels, but on the relevant positions of the classname in the text prompt.

- **Detection** only cares about the proposed objects of the current frame. Therefore, in a unified detection-tracking context, we simply treat all detected objects as newly appeared objects, and the bipartite matching happens between the proposed object queries and the ground truth objects.

- **Tracking**, in addition to localization and classification of objects in the current frame, requires the knowledge of object/track identities across video frames, and faces the challenges of objects disappearing or re-entering the scene. Thanks to the self-attention mechanism in transformers [96] which correlates all components across the entire inputs, data association across video frames could be achieved in a detection/tracking-by-attention fashion [12, 71, 120]. Specifically, the frame-to-frame data association is realized by 1) integrating previous frame’s features into current frame’s transformer encoder, where a temporal feature encoding [100] is used to enable queries to discriminate between features from the previous frame; and 2) adding the previous detected object queries, named track queries, to the initialization of new object queries for the current frame, and together forward into the transformer’s decoder of current frame (Fig. 2). In the transformer decoder, computing self-attention between adjacent frame features as well as between newly initialized object queries and track queries, naturally performs the detection of new objects while avoiding re-detection of already detected/tracked objects [71].

Therefore, the bipartite matching for tracking contains two scenarios. 1) If the objects in the current frame are...
also present in the previous frame, the mapping depends on the ground truth track identities [71]; 2) Otherwise, the mappings to newly-appeared objects or background reduce to the same matching plan as detection [12, 120].

In summary, the bipartite matching loss, for either detection or tracking, is achieved by solving a minimum cost assignment problem [12], resulting in the following combined end-to-end training loss for TrIVD:

$$L = L_{\text{soft}} + L_{\text{contrastive}} + L_{\text{box\_detect}} + L_{\text{box\_track}},$$ (5)

where $L_{\text{soft}}$, $L_{\text{contrastive}}$ are the object category prediction losses (Sec. 3.2), $L_1$ loss and Generalized IoU [12] are used as the box prediction losses for both tracking object boxes ($L_{\text{box\_track}}$) and newly-appeared/non-object boxes ($L_{\text{box\_detect}}$). Since for detection tasks, we treat all proposed objects as new detections, thus $L_{\text{box\_track}} \equiv 0$.

4. Experiments

4.1. Datasets and Metrics

Image and Video Object Detection (OD)

- COCO For image OD, we experiment on the COCO [64] dataset, with 80 annotated class categories in total. All the models are trained on the 118K training images and evaluated on the 5K validation images.
- VID For video OD, we experiment on the large-scale benchmark for video OD, ImageNet VID [84] dataset. It contains 3862 training videos and 555 validation videos. VID has 30 annotated class categories in total, among them 13 categories overlap with those in COCO. We also follow the previous video OD work [23, 97, 104, 106] and include DET [85] dataset in the training set.

- Metrics For image OD, we use the 6 official metrics on average precision (AP) from COCO [64], i.e., AP, AP$_{50}$, AP$_{75}$, AP$_S$, AP$_M$, and AP$_L$. For video OD, AP is used as the evaluation metric following previous work [23, 97, 104, 106].

Multi-object Tracking (MOT)

- MOT17 We train and test our model’s tracking performance on the MOTChallenge benchmark, MOT17 [73]. MOT17 is person-annotated only, and has 7 sequences for train and test sets respectively.
- Metrics Varying metrics are used for evaluating different aspects of MOT performance [7, 71]. We adopt the 7 widely-used metrics [73, 83]: multiple object tracking accuracy (MOTA), identity F1 score (IDF1), mostly tracked (MT), mostly lost (ML), false positive (FP) and false negative (FN), and number of identity switches (IDS). A detailed description on the evaluation metrics is in Appendix A.

4.2. Implementation Details

Training We use MViTv2-s [27, 60] as the backbone, we follow deformable DETR [120] for its end-to-end transformer-based structure, and TrackFormer [71] for its track queries aggregation and augmentations. To make sure we can cover objects in the crowded scenes in MOT17 [73] tracking dataset, we set the number of object queries to $N_{\text{box}} = 500$. For the MViTv2-s backbone, we follow [60] and pre-train the backbone on ImageNet-21K [22] and fine-tune on COCO with 36 epochs. Our training sched-
4.3. Individual-dataset Benchmark Results

We explore our unified model’s performance on COCO [64], VID [84] and MOT17 [73] datasets. To better illustrate the benefits gained from our unified formulations, we explore two training setups for TrIVD. 1) \textbf{TrIVD}_{\text{single}}: the proposed TrIVD model, but trained on individual datasets separately; 2) \textbf{TrIVD}_{\text{multi}}: the proposed TrIVD model, co-trained on all three datasets (COCO [64], VID [84], MOT17 [73]) in a multi-dataset, multi-task fashion.

Tabs. 1-3 compare TrIVD’s performance with the state-of-the-art approaches on all image OD, video OD and MOT tasks. TrIVD achieves state-of-the-art performance across all evaluation metrics on COCO [64], especially on small objects (AP50) (Tab. 1). Comparisons on VID [84] dataset demonstrate the effectiveness of the proposed temporal-aware attention. Table 2. Comparisons between TrIVD and the state-of-the-art video OD approaches on VID validation set. \(N_{\text{frame}}\) refers to the temporal length of corresponding models’ input video clips. Deformable-DETR (Def-DETR) [120] could be viewed as our per-frame detection baseline on video OD: \textbf{TrIVD}_{\text{single}} equals Def-DETR when we switch the MViTv2-s [60] backbone to ResNet-50 [40] and perform frame-by-frame detection for VID, i.e., without temporal-aware attention.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Detector</th>
<th>(N_{\text{frame}})</th>
<th>AP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFF [104]</td>
<td>ResNet-50</td>
<td>Faster-RCNN</td>
<td>10</td>
<td>70.4</td>
</tr>
<tr>
<td>FGFA [103]</td>
<td>ResNet-50</td>
<td>Faster-RCNN</td>
<td>21</td>
<td>74.0</td>
</tr>
<tr>
<td>RDN [23]</td>
<td>ResNet-50</td>
<td>Faster-RCNN</td>
<td>3</td>
<td>76.7</td>
</tr>
<tr>
<td>MEGA [17]</td>
<td>ResNet-50</td>
<td>Faster-RCNN</td>
<td>9</td>
<td>77.3</td>
</tr>
<tr>
<td>TransVOD [106]</td>
<td>ResNet-50</td>
<td>Def-DETR</td>
<td>3</td>
<td>77.7</td>
</tr>
<tr>
<td>Def-DETR [120]</td>
<td>ResNet-50</td>
<td>Def-DETR</td>
<td>1</td>
<td>76.0</td>
</tr>
<tr>
<td>\textbf{TrIVD}_{\text{single}}</td>
<td>MViTv2-s</td>
<td>Def-DETR</td>
<td>3</td>
<td>77.9</td>
</tr>
<tr>
<td>\textbf{TrIVD}_{\text{multi}}</td>
<td>MViTv2-s</td>
<td>Def-DETR</td>
<td>3</td>
<td>\textbf{78.3}</td>
</tr>
</tbody>
</table>

Table 1. Comparisons between TrIVD and the state-of-the-art image OD approaches on COCO 2017 validation set. Deformable-DETR (Def-DETR) [120] could be viewed as our plain baseline on image OD: \textbf{TrIVD}_{\text{single}} equals Def-DETR when we switch the MViTv2-s [60] backbone to ResNet-50 [40].

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Backbone</th>
<th>Detector</th>
<th>MOTA ↑</th>
<th>IDF ↑</th>
<th>MT ↑</th>
<th>ML ↓</th>
<th>FP ↑</th>
<th>FN ↓</th>
<th>IDS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAMNet [19]</td>
<td>COCO</td>
<td>ResNet-101</td>
<td>52.0</td>
<td>48.7</td>
<td>450</td>
<td>787</td>
<td>14118</td>
<td>253616</td>
<td>3072</td>
<td></td>
</tr>
<tr>
<td>MEGA [17]</td>
<td>COCO</td>
<td>ResNet-53</td>
<td>56.4</td>
<td>57.8</td>
<td>523</td>
<td>813</td>
<td>14379</td>
<td>230174</td>
<td>1485</td>
<td></td>
</tr>
<tr>
<td>CenterTrack [116]</td>
<td>DLA [107]</td>
<td>62.5</td>
<td>55.7</td>
<td>560</td>
<td>777</td>
<td>11599</td>
<td>208577</td>
<td>2540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMOH [90]</td>
<td>-</td>
<td>ResNet-101</td>
<td>62.1</td>
<td>62.8</td>
<td>633</td>
<td>759</td>
<td>10951</td>
<td>201195</td>
<td>1897</td>
<td></td>
</tr>
<tr>
<td>TrackFormer [71]</td>
<td>-</td>
<td>ResNet-50</td>
<td>62.3</td>
<td>57.6</td>
<td>688</td>
<td>638</td>
<td>16591</td>
<td>192123</td>
<td>4018</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Comparisons between TrIVD and the state-of-the-art MOT approaches on MOT17 test set (online public detections reported). The 2\textsuperscript{nd} column indicates extra tracking data included in the training (M: Market1501 [114]; C: CUHK03 [59]).
aware attention module in aggregating spatio-temporal features of video inputs (Tab. 2). We also achieve better MOT performance on MOTA, MT, ML as well as FN, without the need for training on additional tracking data as used in [6, 65] (Tab. 3).

Comparisons between the dataset-specific trained model (TrIVDsingle) and the multi-dataset multi-task jointly trained model (TrIVDmulti) further illustrate the effectiveness of multi-task co-training with our proposed unified formulation – TrIVDmulti outperforms TrIVDsingle over image OD, video OD and MOT tasks, especially for MOT where we observe an improvement on MOTA by 3.7% (Tab. 3).

4.4. Cross-dataset Visualization Analysis

In Sec. 4.3, we report the benchmark results of TrIVD on image OD, video OD and MOT tasks. However, with the unified formulation, TrIVD can achieve much more than that. 1) TrIVD is able to perform zero-shot tracking, on objects whose categories do not exist in the tracking training dataset. 2) With our unified classifier via region-text alignment, TrIVD’s detection/tracking vocabulary could be easily scaled up upon co-training with larger datasets, and achieves open-vocabulary capabilities.

Zero-shot Tracking TrIVD’s unified formulation allows us to conduct image OD, video OD and MOT within one model (Fig. 1), and further extend tracking to a wider range of object categories, achieving zero-shot tracking capability. As shown in Fig. 3, designed and trained specifically on MOT17, a person-tracking dataset, a typical tracking model such as TrackFormer [71] can detect and track people identities (Fig. 3, 1st column), but it is not able to track other object categories that are not annotated in MOT17, e.g., cars, birds, pandas. In contrast, with our unified formulation, TrIVD that is co-trained on OD datasets (COCO, VIDD) is now able to borrow its knowledge learned from the detection data and achieve zero-shot tracking on novel objects without training on their tracking annotations. We can therefore track both the people and the cars in the same street view from MOT17 (Fig. 3, 2nd–3rd columns).

We further test TrIVD’s zero-shot tracking ability on videos from VIDD (video OD) dataset, where no ground truth tracking annotation is available (Fig. 1, 2nd–4th columns; Fig. 3, 4th–7th columns). TrIVD successfully detects and tracks the objects with their position changes (e.g., airplanes and pandas in Fig. 1, motorcycles in Fig. 3). TrIVD also handles object disappearing scenarios well (e.g., cars in Fig. 1, airplanes in Fig. 3). Besides objects disappearance, another major challenge for MOT is identifying previously tracked objects that re-enter the scene — we indeed observe some failure cases. In Fig. 3, 6th column, TrIVD successfully re-identifies the bird (pink) when it re-enters the scene, but fails to recognize the same bicycle (Fig. 3, 7th column) under the significant camera view and pose changes, and identifies it as a new bicycle (blue → green).

Open-vocabulary Detection/Tracking In Fig. 1 and Fig. 4, we show TrIVD’s performance on OD, where we detect in images as well as videos that contain blurred or occluded objects (Fig. 1, 5th column; Fig. 4, 2nd column). Furthermore, with our unified cross-dataset classifier by formulating detection/tracking as phrase grounding (Sec. 3.2), we naturally extend the model’s detection ability to the combined annotated object categories of the three co-trained datasets (COCO, VIDD, MOT17). Therefore, in addition to people, TrIVD also detects other objects (e.g., cars, bicycles, motorcycles) in the street videos from the person-annotated only MOT17 dataset (Fig. 4, 3rd column).

Since TrIVD predicts object categories based on region-text alignment instead of class labels, with our grounded-formulated unified classifier, TrIVD’s detection/tracking vocabulary could be further scaled up upon pre-training on larger OD datasets such as Objects365 [86], and on semantic-rich phrase grounding datasets [58, 110], e.g., Flickr30K [80], VG Caption [53]. Our on-going research focuses on TrIVD’s open-vocabulary abilities to achieve detecting/tracking in the wild.

4.5. Ablations

We have already explored the effectiveness of our unified formulation in improving the overall performance by comparing TrIVDsingle (trained on individual datasets) with TrIVDmulti (co-trained with multi-dataset, multi-task learning) in Sec. 4.3. To further demonstrate the importance of the proposed temporal-aware attention module (Sec. 3.1) for video OD tasks, we also conduct ablations on TrIVDsingle regarding the number of video frames used to aggregate for video OD. Without any temporal-aware feature aggregation (Nframe = 1) the model reduces to a frame-by-frame image OD model. As shown in Tab. 4, with only 2 additional reference frames forwarded in temporal feature fusion, we observe a significant improvement on TrIVDsingle’s performance. Overall, a temporal length of 7 for input video clip achieves the best video OD performance, while continuing to increase the number of aggregated frames does not bring obvious gains further. More experimental results are provided in Appendix B.

<table>
<thead>
<tr>
<th>Nframe</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP↑</td>
<td>76.8</td>
<td>77.9</td>
<td>78.6</td>
<td>79.4</td>
<td>79.2</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Table 4. Ablations on the number of frames aggregated in an input video clip for video OD on VIDD [84] dataset (Model: TrIVDsingle).
5. Conclusion

We introduced TRIVD which performs image object detection, video object detection, and multi-object tracking within a unified framework. We unified image and video inputs with spatial-temporal-aware deep feature fusion, and connected image-video detection with tracking via self-attention. Our unified classifier based on region-text alignment naturally extends the detection/tracking vocabulary of TRIVD and enables zero-shot tracking. We hope this work brings deeper insights and reveals greater power of the multi-task learning for image-video object detection and multi-object tracking.

References

age is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3


[27] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 3, 6, 14


[36] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. 2020. 3


Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal


[111] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In CVPR, 2020. 2


A Unified Model for Tracking and Image-Video Detection Has More Power (Appendix)

This section provides: 1) more implementation details and training recipes for TrIVD (Appendix A); 2) additional experimental results (Appendix B).

A. Additional Implementation Details

A.1. Backbone Pre-training
For the MViTv2-s [60] backbone training, we follow the same recipe as in [27, 60, 67]. Specifically, we pre-train MViTv2-s on ImageNet-21K [22] for 300 epochs with batch size of 32. We use the truncated normal distribution initialization [38] and synchronized AdamW [68] optimization, with a base learning rate of $2 \times 10^{-3}$ and a linear warm-up in the first 70 epochs followed by a decayed half-period cosine schedule [95]. We set the weight decay to 0.05. We also use stochastic depth [44] with rate as 0.1. The augmentation strategies are the same as in [27, 60].

A.2. Text Encoder
We follow [47] and use the HuggingFace [101] pre-trained RoBERTa-base [66] as our text encoder. In all the experiments, we use a linear decay with warm-up schedule, increasing linearly to $5 \times 10^{-5}$ during the first 1% of the total number of iterations, then decreasing linearly back to 0 for the rest of the training.

A.3. Single-dataset Training
For the single-dataset training of TrIVD\text{single}, we fine-tune the pre-trained model from Appendix A.1 on the downstream object detection and multi-object tracking tasks respectively.

COCO Image Object Detection For image object detection experiments on COCO [64], we follow the 3× schedule (36 epochs) suggested in [60]. We set the batch size as 2 with an initial learning rate of $10^{-4}$ for deformable DETR encoder-decoder, and $10^{-5}$ for the backbone. The learning schedule for the text encoder is detailed in Appendix A.2.

VID Video Object Detection For video object detection experiments on VID [84], we follow the previous work [23, 97, 104, 106] and include DET [85] dataset in the pre-training stage. Then, we train our model on VID in batch size as 1 for 3 epochs. The initial learning rate for deformable DETR is $5 \times 10^{-4}$, and $5 \times 10^{-5}$ for the backbone. The learning schedule for the text encoder is detailed in Appendix A.2.

MOT17 Multi-object Tracking For multi-object tracking experiments on MOT17 [73], we initialize the model weights from the model trained on COCO [64] as described in Appendix A.3. We train our model on MOT17 [73] with a batch size of 2 for 50 epochs with a learning rate drop to 0.1× after 30 epochs. The initial learning rate for deformable DETR is $5 \times 10^{-4}$, and $5 \times 10^{-5}$ for the backbone. The learning schedule for the text encoder is detailed in Appendix A.2.

A.4. Cross-dataset Co-training
For the joint training of TrIVD\text{multi} on all the three datasets (COCO, VID, MOT), we initialize the model weights from the model trained on COCO [64] as in Appendix A.3. We then co-train our model on the combined dataset of COCO, VID and MOT with a batch size of 1 for 120 epochs with a learning rate drop to $0.1$ after 80 epochs. The initial learning rate for deformable DETR is $10^{-4}$, and $10^{-5}$ for the backbone. The learning schedule for the text encoder is detailed in Appendix A.2.

To handle the different tasks (image object detection, video object detection, or multi-object tracking) for samples from different datasets, we separate samples from the three datasets in each forward pass. To balance the varying scales of different datasets, we randomly select 20,000 video clip samples from VID [84] in each training epoch.

A.5. Track Initialization and Re-identification
New objects appearing in the current frame compared to previous frames are detected by a fixed number of $N_{\text{box}} = 500$ object queries [120], each attends to certain spatial locations in the current frame [120]. Given the object encodings, the deformable-DERT transformer decoder’s self-attention intrinsically avoids duplicate detections [12, 71, 120]. For the tracking purpose, each newly detected objects will also initialize a new track query with its associated object embedding [71]. Track queries then follow the corresponding objects based on their embeddings throughout a video and adapt to the position changes simultaneously. Depends on the objects status in the given video sequences, the number of track queries $N_{\text{track}}$ could change across frames each time when new objects are detected or previously-detected tracks disappear or are occluded. Following [71], we remove a detection/track when its classification confidence score drops below $\sigma_{\text{track}} = 0.4$, or is lower than an IoU threshold $\sigma_{\text{NMS}} = 0.9$ for non-maximum suppression (NMS).

During tracking inference, we use previously proposed track queries for an attention-based re-identification process. We follow [71] and keep previously removed track queries within an optimal inactive patience of $N_{\text{reid}} = 5$ frames, during which the track queries are considered as not active and thus are not used in the object queries initial-
ization of new frames, unless a classification score higher than $\sigma_{\text{reid}} = 0.4$ triggers the re-identification.

**A.6. Track Filtering**

Typical tracking-by-detection methods [16,42,50,51,56,57] perform data association on a bounding box level during tracking evaluations. Yet this strategy is not suitable for tracking-by-attention or point-based methods [71,116]. To achieve a fairer comparison, we follow [71] and perform the track filtering on Intersection over Union (IoU) and initialize tracks with IoU greater than 0.5.

**A.7. Evaluation Metrics for Multi-object Tracking**

We provide brief definitions of the seven evaluation metrics used for MOT comparisons (Sec. 4.3) in the main paper. For more complete analyses on different metrics for multi-object tracking, please also refer to [73,83].

**False Negative (FN)** refers to the number of false negative ground truth bounding boxes that are not covered by any bounding box.

**False Positive (FP)** refers to the number of false positive bounding boxes that do not correspond to any ground truth object.

**Multiple Object Tracking Accuracy (MOTA)** penalizes detection errors ($\text{FN} + \text{FP}$) and fragmentations ($\Phi$) normalized by the total number $N$ of true detections:

$$\text{MOTA} = 1 - \frac{\text{FN} + \text{FP} + \Phi}{N}. \quad (A.6)$$

**IDF1** is defined as the ratio between the correctly identified detections and the average number of ground truth objects and computed detections:

$$\text{IDF1} = \frac{2 \text{IDTP}}{2 \text{IDTP} + \text{IDFP} + \text{IDFN}}, \quad (A.7)$$

where IDP refers to the identification precision. And IDTP, IDFP, IDFN are the true positive, false positive, and false negative of IDP respectively.

**Mostly Tracked (MT)** denotes the number of tracks that are successfully tracked for larger than 80% of its total span.

**Mostly Lost (ML)** denotes the number of tracks that are successfully tracked for less than 20% of its total length.

**Identity Switch (IDS)** counts the number of mismatches of a ground truth object that is originally identified as track $i$ but assigned to another track $j$ ($i \neq j$) in the following frames.

**Figure A.5.** TrIVD’s object detection results on COCO, VID, and MOT17. With multi-dataset co-training, TrIVD detects objects (e.g., traffic lights, bicycles, motorcycles) not annotated by MOT17 (3rd column).

**B. Additional Experimental Results**

**B.1. Zero-shot Multi-object Tracking**

In Fig. A.6 and Fig. A.7, we present more complete zero-shot multi-object tracking results from TrIVD, in addition to Fig. 1 and Fig. 3 in the main paper.

**B.2. Cross-dataset Detection**

We provide more cross-dataset detection results in Fig. A.5 on all the three experimented datasets (COCO [64], VID [84], MOT [73]) in addition to Fig. 1 and Fig. 4 in the main paper.
<table>
<thead>
<tr>
<th>t</th>
<th>Track “airplane”</th>
<th>Track “bird”</th>
<th>Track “giant panda”</th>
<th>Track “car”</th>
</tr>
</thead>
</table>

Figure A.6. Visualizations on TrIVD’s zero-shot multi-class, multi-object tracking performance (Sec. 4.4). All the above videos are from VID [84] where no tracking annotation is available.
Figure A.7. Visualizations on TrIVD’s zero-shot multi-class, multi-object tracking performance (Sec. 4.4). All the above videos are from VID [84] where no tracking annotation is available. In the last column, we show a failure case where the model recognizes the “blue-box” bicycle as a new track (“green-box”) when it re-enters the scene, due to the significant pose and camera view changes.