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

Peirong Liu[1 *] Rui Wang[2] Pengchuan Zhang[2] Omid Poursaeed[2] Yipin Zhou[2]
Xuefei Cao[2] Sreya Dutta Roy[2] Ashish Shah[2] Ser-Nam Lim[2]

[1]UNC-Chapel Hill [2]Meta AI

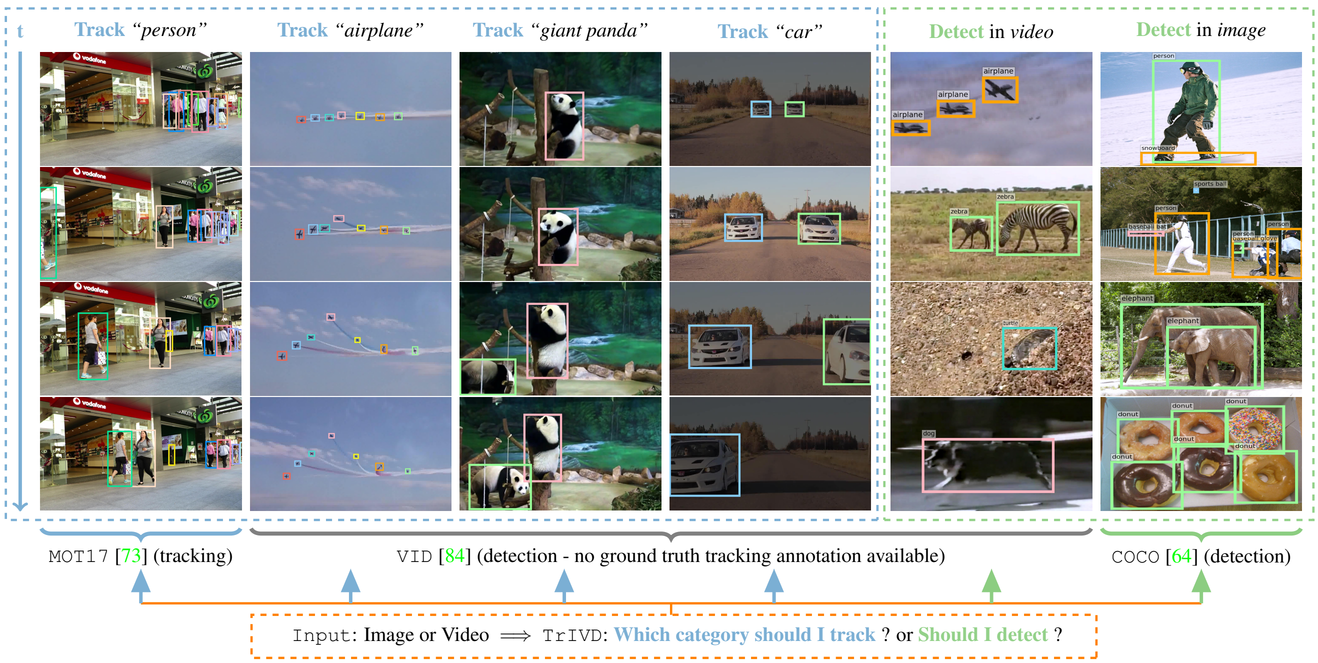

Figure 1. TrIVD enables image/video object detection and multi-object tracking within a single model. With the proposed unified framework, we are uniquely able to conduct *zero-shot* multi-object tracking on objects (airplanes, pandas, etc.) that have *not* appeared in tracking datasets. (Different colors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

Abstract

Objection 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 $B' = B * T$ 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, given 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 Φ 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,

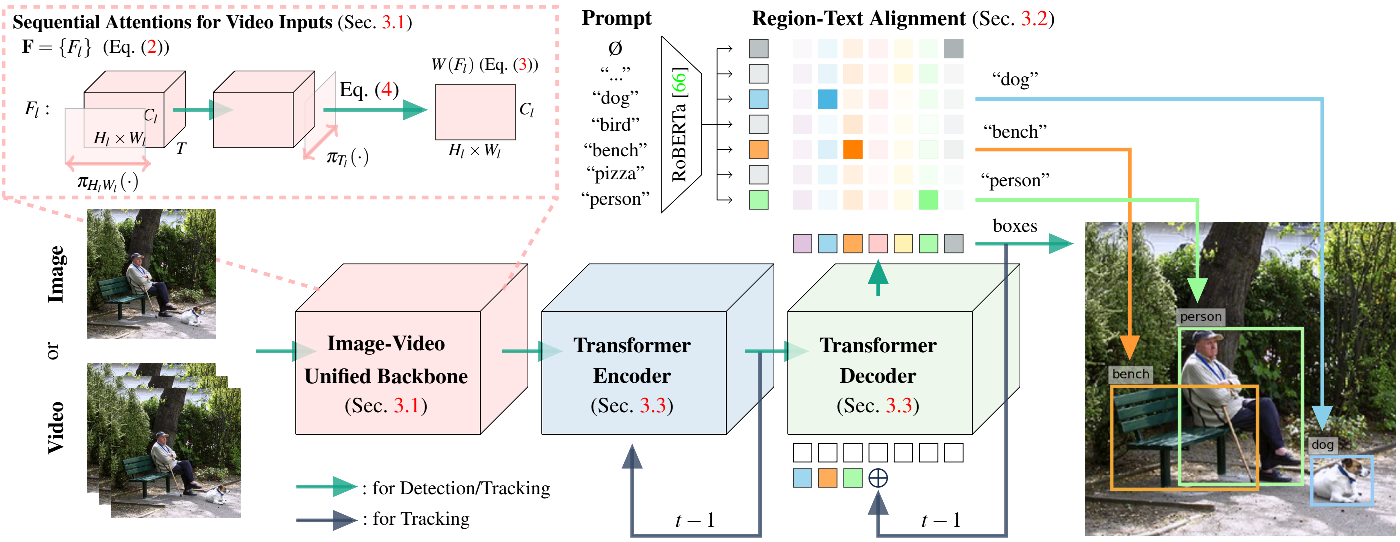

Figure 2. Overview of `TrIVD`'s unified framework. With a unified backbone, `TrIVD` could take both images and videos as inputs, using the proposed temporal-aware attention for spatio-temporal feature fusion of video features (Sec. 3.1). Depending on specific tasks, `TrIVD` performs object detection 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 the detected objects from the previous frame (Sec. 3.3). Formulating the object category prediction as phrase grounding, we determine the object class by assessing the alignment between the proposed regions and the words in an input text prompt (Sec. 3.2).

$$\overline{\mathbf{F}} = \{\overline{F_l} \mid \overline{F_l} \in \mathbb{R}^{B' \times H_l \times W_l \times C_l}, l = 1, ..., L\}, \quad (1)$$

where $H_l \times W_l$ refer to the spatial resolution at scale l, and L denotes the number of spatial scales. Features of video inputs are then re-shaped to their original dimensions:

$$\mathbf{F} = \{F_l | F_l \in \mathbb{R}^{B \times T \times H_l \times W_l \times C_l}, l = 1, ..., L\}, \quad (2)$$

where B, T denote the actual batch size and temporal length of the inputs respectively, with $T \equiv 1$ for image inputs.

Our unified input formulation shares similar spirits with Omnivore [29], however, Omnivore represents both image and video inputs as *videos*, i.e., as batches of *4D* tensors $X \in \mathbb{R}^{B \times T \times H \times W \times C}$, where the temporal length (T) for image inputs are set to 1. This results in 3D operations (e.g., 3D convolutions) and more expensive computations especially for video OD/MOT problems that require high-resolution inputs. Instead, we treat video inputs as batches of *images*, and conduct frame-level feature extraction first. Spatial features extracted in the backbone are then forwarded to our temporal-aware sequential attentions as described below, for spatio-temporal feature fusion.[1]

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

$$W(F_l) = \pi_{T_l}\left(\pi_{H_l W_l}(F_l)\right) \cdot F_l, \quad l = 1, ..., L, \quad (3)$$

where $\pi_{T_l}(\cdot)$, $\pi_{H_l W_l}(\cdot)$ 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 ($\pi_{H_l W_l}$), and apply an additional temporal attention module by dynamically aggregating features across their temporal dimensions:

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

where $f(\cdot)$ is a linear function approximated by a 1×1 convolutional layer and σ 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

[1] In 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.

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 belonged dataset to form a text prompt. For instance, `VID` [84] dataset has the label to classname correspondences,

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

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

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

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 ($\mathscr{L}_{\text{soft}}$) [47], to encourage the predicted token spans to be aligned with the objects' semantic meanings, and 2) the contrastive alignment loss ($\mathscr{L}_{\text{contrastive}}$) [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_{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_{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

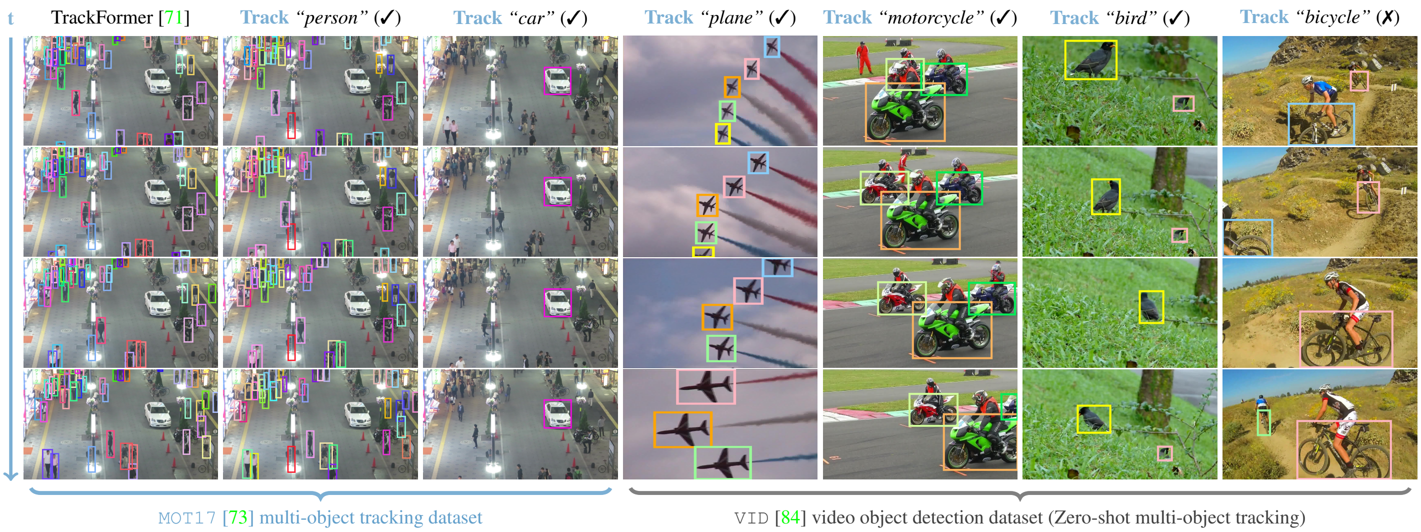

Figure 3. With its unified formulation, `TrIVD` achieves unique capability in zero-shot *multi-class*, *multi-object* tracking (Sec. 4.4). 1st column shows person-tracking results from TrackFormer [71]. `TrIVD` is not only able to track people (2nd column), but also able to track the car in the same scene when assigned which category to track (3rd column). Further, `TrIVD` achieves zero-shot tracking on objects from `VID`, which do *not* exist in the tracking dataset (`MOT17`) for supervised training (4rd – 6th columns). We also show videos with the challenging problem of objects disappearing and re-entering, where we observe some failure cases (6th column: tracked (✓); 7th column: failed (✗)).

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`:

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

where $\mathscr{L}_{\text{soft}}$, $\mathscr{L}_{\text{contrastive}}$ 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 ($\mathscr{L}_{\text{box\_track}}$) and newly-appeared/non-object boxes ($\mathscr{L}_{\text{box\_detect}}$). Since for detection tasks, we treat all proposed objects as new detections, thus $\mathscr{L}_{\text{box\_track}} \equiv 0$.

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-

Method	Backbone	Detector	AP ↑	AP_{50} ↑	AP_{75} ↑	AP_S ↑	AP_M ↑	AP_L ↑
Faster-RCNN [30]	ResNet-50	Faster-RCNN	42.0	62.1	45.5	26.6	45.4	53.4
DETR [12]	ResNet-50	DETR	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5 [12]	ResNet-50	DETR	43.3	63.1	45.9	22.5	47.3	61.1
Def-DETR [120]	ResNet-50	Def-DETR	43.8	62.6	47.7	26.4	47.1	58.0
Def-DETR$^{\text{box-refine}}$ [120]	ResNet-50	Def-DETR	45.4	64.7	49.0	26.8	48.3	61.7
TrIVD$_{\text{single}}$	MViTv2-s	Def-DETR	46.2	65.1	48.9	**27.9**	48.9	61.6
TrIVD$_{\text{multi}}$	MViTv2-s	Def-DETR	**46.5**	**65.7**	**49.3**	27.5	**48.9**	**61.9**

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: `TrIVD`$_{\text{single}}$ equals Def-DETR when we switch the MViTv2-s [60] backbone to ResNet-50 [40].

Method	Backbone	Detector	N_{frame}	AP ↑
DFF [104]	ResNet-50	Faster-RCNN	10	70.4
FGFA [103]	ResNet-50	Faster-RCNN	21	74.0
RDN [23]	ResNet-50	Faster-RCNN	3	76.7
MEGA [17]	ResNet-50	Faster-RCNN	9	77.3
TransVOD [106]	ResNet-50	Def-DETR	3	77.7
Def-DETR [120]	ResNet-50	Def-DETR	1	76.0
TrIVD$_{\text{single}}$	MViTv2-s	Def-DETR	3	77.9
TrIVD$_{\text{multi}}$	MViTv2-s	Def-DETR	3	**78.3**

Table 2. Comparisons between `TrIVD` and the state-of-the-art video OD approaches on `VID` validation set. $\mathbf{N}_{\text{frame}}$ 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: `TrIVD`$_{\text{single}}$ reduces to 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.

ules follow [120], and we set the batch size as 2 with initial learning rates of 0.0001 for deformable DETR encoder-decoder, and 0.00001 for the backbone. For the language model, we follow [47] and use the HuggingFace [101] pre-trained RoBERTa-base [66] as our text encoder. We use a linear decay with warm-up schedule, increasing linearly to 0.00005 during the first 1% of the total number of steps, then decreasing linearly back to 0 for the rest of the training. For more details on training recipes and hyperparameters, please refer to Appendix A.

Track Re-identification During tracking inference, we use previously proposed track queries for an attention-based re-identification process. For a fair comparison, we follow [71] and keep previously removed track queries within a tolerance 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 initialization of new frames, unless a classification score higher than $\sigma_{\text{reid}} = 0.4$ triggers the re-identification. More track filtering and re-identification details are provided in Appendix A.

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 for-

Method	Data	Backbone	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDS ↓
FAMNet [18]	-	ResNet-101	52.0	48.7	450	787	14138	253616	3072
Tracktor++ [6]	M & C	ResNet-101	56.3	55.1	498	831	**8866**	235449	1987
GSM [65]	M & C	ResNet-34	56.4	57.8	523	813	14379	230174	**1485**
CenterTrack [116]	-	DLA [107]	60.5	55.7	580	777	11599	208577	2540
TMOH [90]	-	ResNet-101	62.1	**62.8**	633	739	10951	201195	1897
TrackFormer [71]	-	ResNet-50	62.3	57.6	688	638	16591	192123	4018
TrIVD$_{\text{single}}$	-	MViTv2-s	62.5	58.1	671	613	15896	190325	4072
TrIVD$_{\text{multi}}$	-	MViTv2-s	**64.8**	60.1	**724**	**598**	15332	**187232**	3967

Table 3. Comparisons between `TrIVD` and the state-of-the-art MOT approaches on `MOT17` test set (online public detections results reported). The 2nd column indicates extra tracking data included in the training (M: Market1501 [114]; C: CUHK03 [59]).

`COCO` [64] `VID` [84] `MOT17` [73]

Figure 4. `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).

mulation, we explore two training setups for `TrIVD`. 1) `TrIVD`$_{\text{single}}$: the proposed `TrIVD` model, but trained on *individual* datasets separately; 2) `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 (AP_S) (Tab. 1). Comparisons on `VID` [84] dataset demonstrate the effectiveness of the proposed temporal-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 ($\texttt{TrIVD}_{\texttt{single}}$) and the multi-dataset multi-task jointly trained model ($\texttt{TrIVD}_{\texttt{multi}}$) further illustrate the effectiveness of multi-task co-training with our proposed unified formulation – $\texttt{TrIVD}_{\texttt{multi}}$ outperforms $\texttt{TrIVD}_{\texttt{single}}$ 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, 1$^{\text{st}}$ 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`, `VID`) 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, 2$^{\text{nd}}$–3$^{\text{rd}}$ columns).

We further test `TrIVD`'s zero-shot tracking ability on videos from `VID` (video OD) dataset, where no ground truth tracking annotation is available (Fig. 1, 2$^{\text{nd}}$–4$^{\text{th}}$ columns; Fig. 3, 4$^{\text{th}}$–7$^{\text{th}}$ 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, 6$^{\text{th}}$ column, `TrIVD` successfully re-identifies the bird (pink) when it re-enters the scene, but fails to recognize the same bicycle (Fig. 3, 7$^{\text{th}}$ column) under the significant camera view and pose changes, and identifies it as a new bicycle (blue → green).

$\mathbf{N}_{\text{frame}}$	**1**	**3**	**5**	**7**	**9**	**11**
AP ↑	76.8	77.9	78.6	**79.4**	79.2	79.3

Table 4. Ablations on the number of frames aggregated in an input video clip for video OD on `VID` [84] dataset (Model: $\texttt{TrIVD}_{\texttt{single}}$).

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, 5$^{\text{th}}$ column; Fig. 4, 2$^{\text{nd}}$ 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`, `VID`, `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, 3$^{\text{rd}}$ column).

Since `TrIVD` predicts object categories based on region-text alignment instead of class labels, with our grounding-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 $\texttt{TrIVD}_{\texttt{single}}$ (trained on individual datasets) with $\texttt{TrIVD}_{\texttt{multi}}$ (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 $\texttt{TrIVD}_{\texttt{single}}$ regarding the number of video frames used to aggregate for video OD. Without any temporal-aware feature aggregation ($\mathbf{N}_{\text{frame}} = 1$) 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 $\texttt{TrIVD}_{\texttt{single}}$'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.

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

[1] Ana C. Murillo Alberto Sabater, Luis Montesano. Robust and efficient post-processing for video object detection. In *International Conference of Intelligent Robots and Systems (IROS)*, 2020. 3

[2] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 3

[3] Relja Arandjelović and Andrew Zisserman. Objects that sound. *ArXiv*, abs/1712.06651, 2018. 3

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[5] Hatem Belhassen, Heng Zhang, Virginie Fresse, and El-Bay Bourennane. Improving video object detection by seq-bbox matching. In *VISIGRAPP*, 2019. 3

[6] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951, 2019. 2, 3, 7, 8

[7] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[8] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. *ArXiv*, abs/1803.05549, 2018. 3

[9] Guillem Bras'o and Laura Leal-Taix'e. Learning a neural solver for multiple object tracking. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6246–6256, 2020. 2, 3

[10] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021. 2

[11] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1971–1980, 2019. 2

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 5, 6, 7, 14

[13] Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 2004. 3

[14] Lluís Castrejón, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2949, 2016. 3

[15] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In *CVPR*, 2018. 3

[16] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018. 2, 3, 15

[17] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. 2020. 3, 7

[18] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6171–6180, 2019. 7

[19] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7373–7382, June 2021. 4

[20] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian D. Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *Int. J. Comput. Vis.*, 129:845–881, 2021. 2, 3

[21] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6677–6686, 2019. 3

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6, 14

[23] Jiajun Deng, Yingwei Pan, Ting Yao, Wen gang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7022–7031, 2019. 3, 6, 7, 14

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An im-

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

[25] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 3

[26] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015. 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

[28] Golnaz Ghiasi, Barret Zoph, Ekin Dogus Cubuk, Quoc V. Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8836–8845, 2021. 3

[29] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022. 4, 5

[30] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2, 7

[31] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 2

[32] Yunchao Gong, Liwei Wang, Micah Hodosh, J. Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014. 3

[33] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 2, 3

[34] Chaoxu Guo, Bin Fan, Jie Gu, Q. Zhang, Shiming Xiang, Véronique Prinet, and Chunhong Pan. Progressive sparse local attention for video object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3908–3917, 2019. 3

[35] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. *Exploiting Better Feature Aggregation for Video Object Detection*, page 1469–1477. Association for Computing Machinery, 2020. 3

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

[37] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Humphrey Shi, Jianan Li, Shuicheng Yan, and Thomas S. Huang. Seq-nms for video object detection. *ArXiv*, abs/1602.08465, 2016. 3

[38] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. In *NeurIPS*, 2018. 14

[39] Fei He, Naiyu Gao, Qiaozhe Li, Senyao Du, Xin Zhao, and Kaiqi Huang. Temporal context enhanced feature aggregation for video object detection. In *AAAI*, 2020. 3

[40] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 7

[41] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021. 2, 3

[42] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. Improvements to frank-wolfe optimization for multi-detector multi-object tracking. *ArXiv*, abs/1705.08314, 2017. 2, 3, 15

[43] Gabriel Huang, Issam Laradji, David Vazquez, Simon Lacoste-Julien, and Pau Rodriguez. A survey of self-supervised and few-shot object detection. In *arXiv preprint arXiv:2110.14711*, 2021. 2

[44] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 14

[45] Zhengkai Jiang, Peng Gao, Chaoxu Guo, Qian Zhang, Shiming Xiang, and Chunhong Pan. Video object detection with locally-weighted deformable neighbors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8529–8536, Jul. 2019. 3

[46] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *European Conference on Computer Vision*, 2020. 3

[47] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 2, 3, 5, 7, 14

[48] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, and Wanli Ouyang. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2018. 3

[49] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017. 3

[50] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:140–153, 2020. 2, 3, 15

[51] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4696–4704, 2015. 2, 3, 15

[52] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5454–5463, 2017. 3

[53] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 8

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. 2

[55] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128(3):642–656, aug 2019. 2

[56] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 418–425, 2016. 2, 3, 15

[57] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 120–127, 2011. 2, 3, 15

[58] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2, 3, 5, 8

[59] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person reidentification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 7

[60] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 3, 6, 7, 14

[61] Lijian Lin, Haosheng Chen, Honglun Zhang, Jun Liang, Yu Li, Ying Shan, and Hanzi Wang. Dual semantic fusion network for video object detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 3

[62] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 42(02):318–327, feb 2020. 2

[63] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020. 5

[64] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1, 6, 7, 14, 15

[65] Qiankun Liu, Q. Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *IJCAI*, 2020. 2, 3, 7, 8

[66] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 4, 7, 14

[67] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 14

[68] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 14

[69] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[70] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1860, 2019. 3

[71] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3, 5, 6, 7, 8, 14, 15

[72] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020. 3

[73] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *ArXiv*, abs/1603.00831, 2016. 1, 6, 7, 14, 15

[74] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *arXiv preprint arXiv:2205.06230*, 2022. 2, 3

[75] Ishan Misra, Abhinav Shrivastava, Abhinav Kumar Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003, 2016. 3

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

agreement. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12470–12481, 2021. 3

[77] Pedro Miguel Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12929–12940, 2021. 3

[78] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 3

[79] Mandela Patrick, Yuki M. Asano, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *ArXiv*, abs/2003.04298, 2020. 3

[80] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015. 8

[81] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 2

[82] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2

[83] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016. 6, 15

[84] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 5, 6, 7, 8, 14, 15, 16, 17

[85] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, dec 2015. 6, 14

[86] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. 8

[87] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 3

[88] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 3

[89] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 2

[90] Daniel S. Stadler and Jürgen Beyerer. Improving multiple pedestrian tracking by track management and occlusion handling. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10953–10962, 2021. 7

[91] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Martin Robertson. Mamba: Multi-level aggregation via memory bank for video object detection. In *AAAI*, 2021. 3

[92] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14454–14463, June 2021. 2

[93] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 2

[94] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A simple and strong anchor-free object detector. 2021. 2

[95] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 14

[96] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3, 5

[97] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *ECCV*, 2018. 6, 14

[98] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

[99] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[100] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *2021*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8737–8746, 2021. 5

[101] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 7, 14

[102] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. *ICCV 2019*, 2019. 3

[103] Jifeng Dai Lu Yuan Yichen Wei Xizhou Zhu, Yujie Wang. Flow-guided feature aggregation for video object detection. 2017. 2, 3, 7

[104] Jifeng Dai Lu Yuan Yichen Wei Xizhou Zhu, Yuwen Xiong. Deep feature flow for video recognition. 2017. 2, 3, 6, 7, 14

[105] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10448–10457, October 2021. 2, 3

[106] Chun-Han Yao, Chen Fang, Xiaohui Shen, Yangyue Wan, and Ming-Hsuan Yang. Video object detection via object-level temporal aggregation. In *ECCV*, 2020. 3, 6, 7, 14

[107] Fisher Yu, Dequan Wang, and Trevor Darrell. Deep layer aggregation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018. 7

[108] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *arXiv preprint arXiv:2203.11876*, 2022. 2, 3

[109] Fangao Zeng, Bin Dong, Tiancai Wang, Cheng Chen, X. Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022. 2, 3

[110] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *ArXiv*, abs/2206.05836, 2022. 2, 3, 5, 8

[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

[112] Yifu Zhang, Pei Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 2, 3

[113] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 3

[114] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 7

[115] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *arXiv preprint arXiv:2201.02605*, 2021. 2, 3, 5

[116] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ArXiv*, abs/2004.01177, 2020. 2, 3, 7, 15

[117] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2

[118] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 2

[119] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[120] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 5, 6, 7, 14

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 2×10^{-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 5×10^{-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 $\texttt{TrIVD}_{\texttt{single}}$, 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` [84] with batch size as 1 for 3 epochs. The initial learning rate for deformable DETR is 5×10^{-4}, and 5×10^{-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 5×10^{-4}, and 5×10^{-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 $\texttt{TrIVD}_{\texttt{multi}}$ 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-DETR 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_{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 initialization 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 (FN + FP) and fragmentations (Φ) normalized by the total number N of true detections:

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FN} + \mathrm{FP} + \Phi}{N}. \tag{A.6}$$

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

$$\mathrm{IDF1} = \frac{2\,\mathrm{IDTP}}{2\,\mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}, \tag{A.7}$$

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.

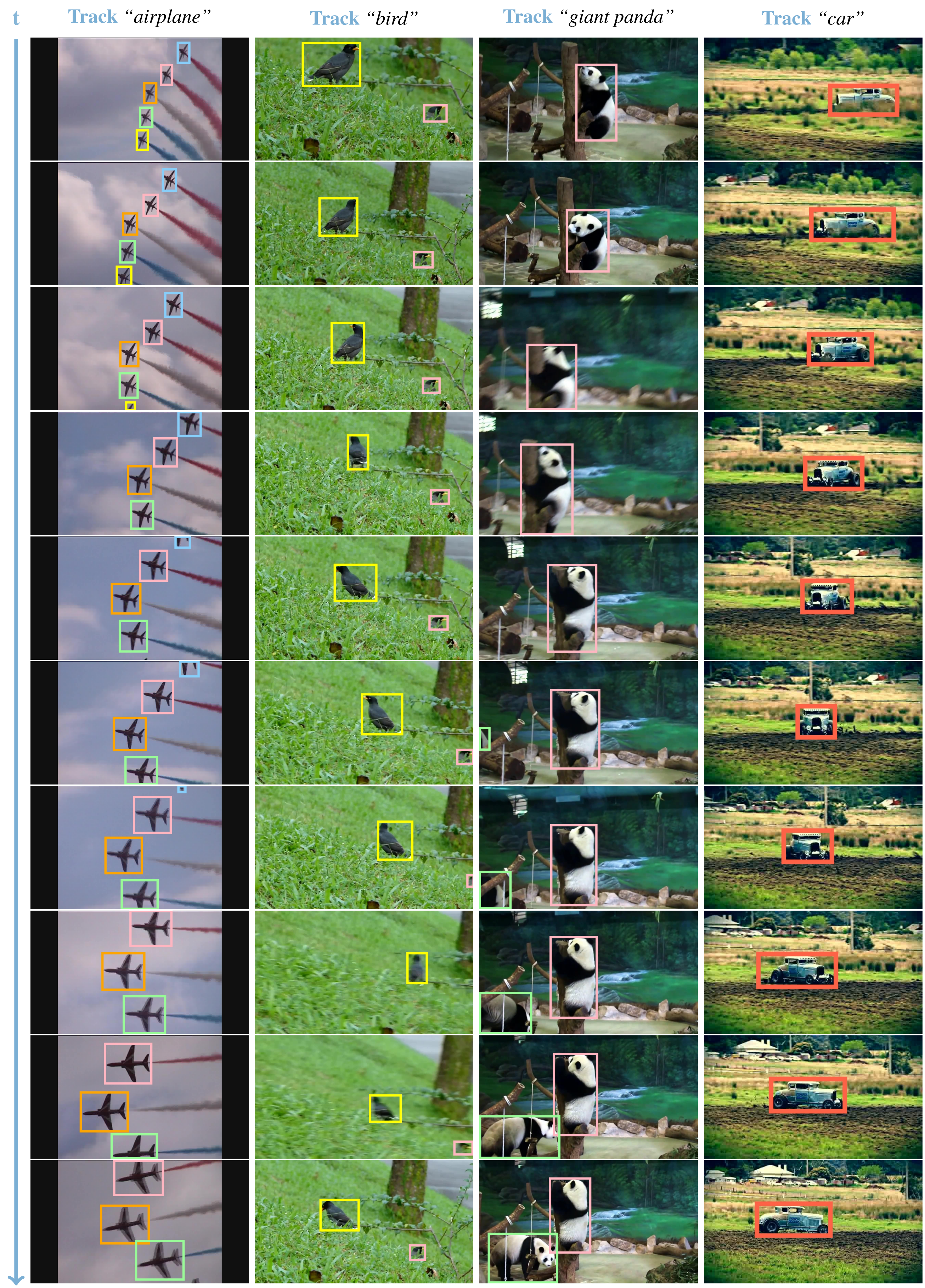

Figure A.6. Visualizations on `TrIVD`'s *zero-shot* multi-class, multi-object tracking performance (Sec. 4.4). All the above videos are from `VID` [84] where no tracking annotation is available.

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.