DAMO-StreamNet: Optimizing Streaming Perception in Autonomous Driving

Jun-Yan He[1] , **Zhi-Qi Cheng**[2] , **Chenyang Li**[1] , **Wangmeng Xiang**[1] ,
Binghui Chen[1] , **Bin Luo**[1] , **Yifeng Geng**[1] , **Xuansong Xie**[1]
[1]DAMO Academy, Alibaba Group
[2]Carnegie Mellon University
{leyuan.hjy, wangmeng.xwm, luwu.lb, cangyu.gyf}@alibaba-inc.com, zhiqic@cs.cmu.edu,
lichenyang.scut@foxmail.com, chenbinghui@bupt.cn, xingtong.xxs@taobao.com

Abstract

In the realm of autonomous driving, real-time perception or streaming perception remains under-explored. This research introduces DAMO-StreamNet, a novel framework that merges the cutting-edge elements of the YOLO series with a detailed examination of spatial and temporal perception techniques. DAMO-StreamNet's main inventions include: (1) a robust neck structure employing deformable convolution, bolstering receptive field and feature alignment capabilities; (2) a dual-branch structure synthesizing short-path semantic features and long-path temporal features, enhancing the accuracy of motion state prediction; (3) logits-level distillation facilitating efficient optimization, which aligns the logits of teacher and student networks in semantic space; and (4) a real-time prediction mechanism that updates the features of support frames with the current frame, providing smooth streaming perception during inference. Our testing shows that DAMO-StreamNet surpasses current state-of-the-art methodologies, achieving 37.8% (normal size (600, 960)) and 43.3% (large size (1200, 1920)) sAP without requiring additional data. This study not only establishes a new standard for real-time perception but also offers valuable insights for future research. The source code is at https://github.com/zhiqic/DAMO-StreamNet.

1 Introduction

The rapid development of autonomous vehicles necessitates robust and efficient traffic environment perception systems. Crucial to this is streaming perception, which concurrently detects and tracks objects in a video stream, and directly influences autonomous driving decisions. Challenges, however, arise from the swiftly fluctuating scales of traffic objects due to vehicle motion, leading to conflicts in the receptive field when identifying both large and small objects. Furthermore, real-time perception is a complex issue largely reliant on motion consistency context and historical data. The two primary hurdles in real-time perception are: (1) the adaptive management of quickly shifting object scales, and (2) the accurate and efficient learning of long-term motion consistency.

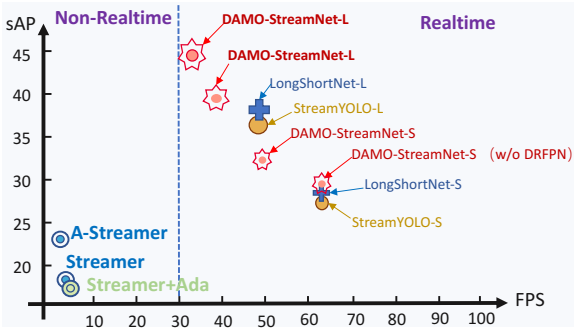
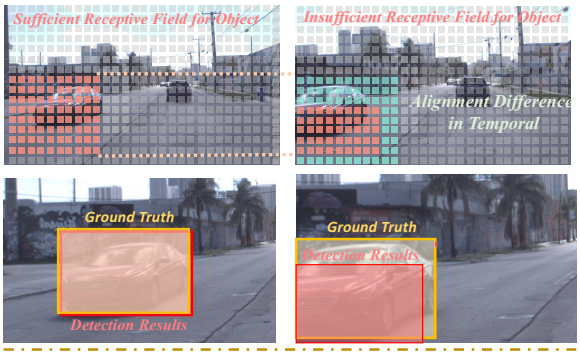


Figure 1: Performance comparisons of streaming perception task, showcasing the balance between accuracy and speed achieved by our proposed method, DAMO-StreamNet, which sets a new state-of-the-art benchmark.

Despite previous research on temporal aggregation techniques [Wang *et al.*, 2018; Chen *et al.*, 2018; Lin *et al.*, 2020; Sun *et al.*, 2021; Huang *et al.*, 2022] has primarily focused on offline settings and is unsuitable for online real-time perception. Furthermore, enhancing the base detector has not been thoroughly investigated in the context of real-time perception. To address these limitations, we propose DAMO-StreamNet, a practical real-time perception pipeline that improves the model in four key aspects:

1. *To augment the performance of the base detector*, we introduce an effective feature aggregation scheme named Dynamic Receptive Field FPN. Leveraging connections and deformable convolution networks, this scheme mitigates receptive field conflicts and strengthens feature alignment capacity. We also implement a state-of-the-art detection technique known as Re-parameterization to boost the network's performance without adding extra inference costs. *These improvements result in superior detection accuracy and quicker inference times.*
2. *To capture long-term spatial-temporal correlations*, we construct a dual-path structure temporal fusion module. *Utilizing a two-stream architecture, this module separates spatial and temporal information, enabling the precise and efficient capture of long-term correlations.*



Resolution	sAP	sAP_{50}	sAP_{75}	sAP_s	sAP_m	sAP_l
600x960	37.1	57.8	37.7	15.2	37.3	63.8
1280x1920	42.7 +5.6	65.4 +7.6	45.0 +7.3	23.9 +8.7	44.8 +7.5	61.7 -2.1

Figure 2: Impact of Receptive Field on Streaming Perception: Inadequate receptive field coverage, as illustrated in the upper and middle regions, leads to unsuccessful predictions. This finding underscores the decrease in performance for large-scale objects with high-resolution input, attributed to limited receptive field coverage.

3. *To address the complexities of learning long-term motion consistency*, we devise an Asymmetric Knowledge Distillation (AK-Distillation) framework. This framework applies a teacher-student learning strategy, wherein student networks are supervised by transferring the generalized knowledge captured by large-scale teacher networks. *This method enforces long-term motion consistency of the feature representations between the teacher-student pair, leading to improved performance.*

4. *To meet the demand for real-time forecasting*, we update the support frame features with the current frame before the subsequent prediction in the inference phase. Furthermore, the support frame features are updated by the current frame in preparation for the next prediction in the inference phase to fulfill the real-time forecasting requirement. *This process enables the pipeline to handle real-time streaming perception and make timely predictions.*

In a nutshell, DAMO-StreamNet presents a cutting-edge solution for real-time perception in autonomous driving. We also introduce a novel evaluation metric, the K-Step Streaming Metric, which takes into account the temporal interval to assess real-time perception. Our experiments demonstrate that DAMO-StreamNet surpasses existing SOTA methods, achieving 37.8% (normal size (600, 960)) and 43.3% (large size (1200, 1920)) sAP without utilizing any extra data. Our work not only sets a new standard for real-time perception but also contributes meaningful insights for future research in this field. Furthermore, DAMO-StreamNet can be adapted to a variety of autonomous systems, such as drones and robots, to provide accurate and real-time environmental perception, thereby enhancing their safety and efficiency.

2 Related Work

2.1 Image Object Detection

State-of-the-art Detectors. The field of image object detection has seen significant progress in recent years due to the development of advanced detectors [Ge *et al.*, 2021b; Wang *et al.*, 2022], with techniques focusing on backbone design [Wang *et al.*, 2021; Ding *et al.*, 2021a; Ding *et al.*, 2021b; Ding *et al.*, 2019; Vasu *et al.*, 2022], feature aggregation [Lin *et al.*, 2017; Ghiasi *et al.*, 2019; Jiang *et al.*, 2022; Tan *et al.*, 2020; Cheng *et al.*, 2022; Tu *et al.*, 2023; Cheng *et al.*, 2017a; Cheng *et al.*, 2019b], and label assignment [Ge *et al.*, 2021a; Kim and Lee, 2020; Carion *et al.*, 2020].

Feature Aggregation. Feature aggregation, especially with FPN [Lin *et al.*, 2017] and PAFPN [Liu *et al.*, 2018], plays a key role in object detection. More recently, the Neural Architecture Search (NAS) methodology has been incorporated into this area [Ghiasi *et al.*, 2019; Cheng *et al.*, 2018; Huang *et al.*, 2018]. GiraffeDet [Jiang *et al.*, 2022; Chen *et al.*, 2023; Zhou *et al.*, 2022] further innovates by using a lightweight backbone and a heavy neck for feature learning.

2.2 Video Object Detection

Temporal Learning. Temporal learning often involves feature aggregation across nearby frames [Wang *et al.*, 2018; Chen *et al.*, 2018; Lin *et al.*, 2020; Sun *et al.*, 2021; Lan *et al.*, 2022; Cheng *et al.*, 2017a]. This has been implemented in DeepFlow [Zhu *et al.*, 2017b] and FGFA [Zhu *et al.*, 2017a] through optic flow, and in MANet [Wang *et al.*, 2018] through pixel-level calibration.

Temporal Linking. Despite the success of temporal learning, video object detection often requires complex temporal modeling components, such as optical flow models [Zhu *et al.*, 2017b], recurrent neural networks [Lin *et al.*, 2020; He *et al.*, 2021], and relation networks [Gao *et al.*, 2021; Cheng *et al.*, 2017b]. Simpler alternatives include temporal linking modules like Seq-NMS [Han *et al.*, 2016], Tubelet rescoring [Kang *et al.*, 2016], and Seq-Bbox Matching [Belhassen *et al.*, 2019; Lan *et al.*, 2022].

2.3 Knowledge Distillation

Knowledge distillation [Hinton *et al.*, 2015] aims to transfer feature representation from a teacher network to a student network. This approach has been adapted in various ways, such as with intermediate-sized teacher-assistant networks [Mirzadeh *et al.*, 2020] and hint learning [Chen *et al.*, 2017]. Other efforts have focused on leveraging different intermediate representations [Heo *et al.*, 2019; Chen *et al.*, 2021; Cheng *et al.*, 2018; Cheng *et al.*, 2019b] or learning data sample or layer relations [Yao *et al.*, 2021; Liu *et al.*, 2020; Yang *et al.*, 2022b; Cheng *et al.*, 2019a]. DAMO-StreamNet is the first work to use knowledge distillation for the streaming perception task, employing the knowledge distillation module to enhance the accuracy of "predicting the next frame" [Yang *et al.*, 2022a] by mirroring the features of the "next frame."

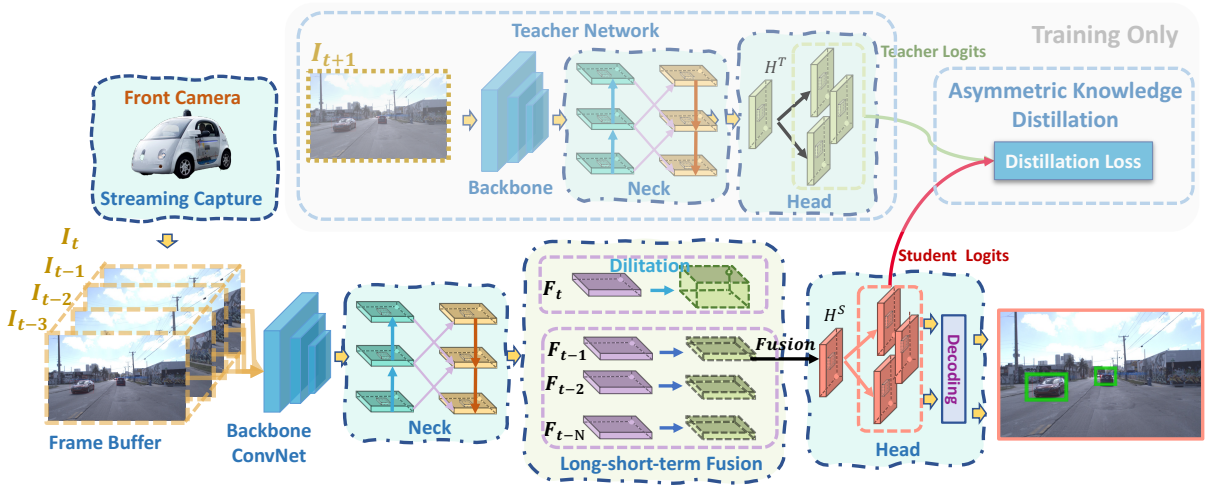


Figure 3: An overview of the proposed DAMO-StreamNet framework. The upper part, obscured by the white mask, contains the teacher network and the Asymmetric Knowledge Distillation module, which are utilized exclusively during the training phase. The lower part represents the student network, featuring the backbone, neck, long-short-term fusion module, and head for efficient streaming perception.

2.4 Streaming Perception

Streaming perception is a relatively new field with limited research focus. Existing methods [Li *et al.*, 2020] are based on object detection and use temporal modeling techniques to improve performance. However, state-of-the-art work such as StreamYOLO [Yang *et al.*, 2022a] does not fully utilize the semantics and motion in video streams. Meanwhile, other recent efforts [Yang *et al.*, 2022a; Li *et al.*, 2022] build on the YOLOX-based detector. Our work with DAMO-StreamNet addresses these issues by re-engineering the base detector and integrating feature aggregation and knowledge distillation. As a result, our method presents a more comprehensive solution for the streaming perception task, outperforming existing methods and setting a new standard for future research.

3 DAMO-StreamNet

The overall framework is illustrated in Fig. 3. Initially, a video frame sequence passes through DAMO-StreamNet to extract spatiotemporal features and generate the final output feature. Subsequently, the Asymmetric Knowledge Distillation module (AK-Distillation) takes the output logit features of the teacher and student networks as inputs, transferring the semantics and spatial position of the future frame extracted by the teacher to the student network.

Given a video frame sequence $\mathcal{S} = \{I_t, \dots I_{t-N\delta t}\}$, where N and δt represent the number and step size of the frame sequence, respectively. DAMO-StreamNet can be defined as,

$$\mathcal{T} = \mathcal{F}(\mathcal{S}, W),$$

where W denotes the network weights, and $\mathcal{T}$ represents the collection of final output feature maps. $\mathcal{T}$ can be further decoded using $Decode(\mathcal{T})$ to obtain the result $\mathcal{R}$, which includes the score, category, and location of the objects.

In the training phase, the student network can be represented as,

$$\mathcal{T}_{stu} = \mathcal{F}_{stu}(\mathcal{S}, W_{stu}).$$

Besides the student network, the teacher network takes the $t+1$ frame as input to generate the future result, represented by,

$$\mathcal{T}_{tea} = \mathcal{F}_{tea}(I_{t+1}, W_{tea}),$$

where W_{stu} and W_{tea} denote the weights of the student and teacher networks, respectively. Then, AK-Distillation leverages $\mathcal{T}_{stu}$ and $\mathcal{T}_{tea}$ as inputs to perform knowledge distillation AKDM$(\mathcal{T}_{stu}, \mathcal{T}_{tea})$. More details are elaborated in the following subsections.

3.1 Network Architecture

The network is composed of three elements: the backbone, neck, and head. It can be formulated as,

$$\mathcal{T} = \mathcal{F}(\mathcal{S}, W) = \mathcal{G}_h(\mathcal{G}_n(\mathcal{G}_b(\mathcal{S}, W_b), W_n), W_h),$$

where $\mathcal{G}_b$, $\mathcal{G}_n$, and $\mathcal{G}_h$ stand for the backbone, neck, and head components respectively, while W_b, W_n, and W_h symbolize their corresponding weights. Previous studies [Jiang *et al.*, 2022] highlighted the neck structure's critical role in feature fusion and representation learning for detection tasks. Consequently, we introduce the Dynamic Receptive Field FPN (DRFPN), which employs a learnable receptive field approach for enhanced feature fusion. To benchmark against the current state-of-the-art (SOTA), we apply the same settings for $\mathcal{G}_n$, $\mathcal{G}_h$, and StreamYOLO [Yang *et al.*, 2022a], leveraging CSPDarknet-53 [Ge *et al.*, 2021b] and TALHead [Yang *et al.*, 2022a] to build the network. Given the proven efficacy of long-term temporal information by the existing LongShortNet [Li *et al.*, 2022], we also integrate a dual-path architectural module for spatial-temporal feature extraction.

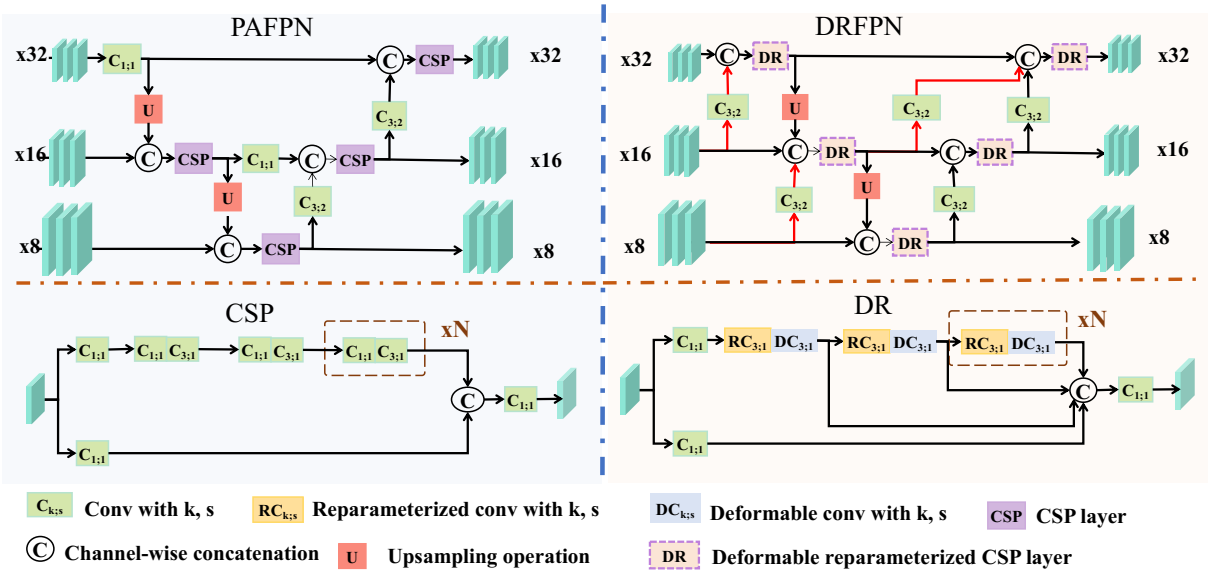


Figure 4: A comprehensive comparison between PAFPN and our proposed DRFPN, both constructed using the base block CSP and DR layer. The notation "Conv with k, s" represents a convolution layer with kernel size 'k' and stride 's'.

Dynamic Receptive Field FPN. Recent object detection studies, including StreamYOLO [Yang *et al.*, 2022a] and LongShortNet [Li *et al.*, 2022], have utilized YOLOX as their fundamental detector. YOLOX's limitation is its fixed spatial receptive field that cannot synchronize features temporally, thus impacting its performance. To address this, we propose the Dynamic Receptive Field FPN (DRFPN) with a learnable receptive field strategy and an optimized fusion mechanism.

Specifically, Fig.4 contrasts PAFPN and DRFPN. PAFPN employs sequential top-down and bottom-up fusion operations to amplify feature representation. However, conventional convolution with a static kernel size fails to align features effectively. As a solution, we amalgamate the DRM module and Bottom-up Auxiliary Connect (BuAC) with PAFPN to create DRFPN. We introduce three notable modifications compared to PAFPN's CSP module (Fig.4):(1) We integrate deformable convolution layers into the DRFPN module to provide the network with learnable receptive fields;(2) To enhance feature representation, we adopt re-parameterized convolutional layers [Ding *et al.*, 2021b];(3) ELAN [Wang *et al.*, 2022] and Bottom-up Auxiliary Connect bridge the semantic gap between low and high-level features, ensuring effective detection of objects at diverse scales.

Dual-Path Architecture. The existing StreamYOLO [Yang *et al.*, 2022a] relies on a single historical frame in conjunction with the current frame to learn short-term motion consistency. While this suffices for ideal uniform linear motion, it falls short in handling complex motion, such as non-uniform motion (e.g., accelerating vehicles), non-linear motion (e.g., rotation of objects or camera), and scene occlusions (e.g., billboard or oncoming car occlusion).

To remedy this, we integrate the dual-path architecture [Li *et al.*, 2022] with a reimagined base detector, enabling the capture of long-term temporal motion while calibrating it with short-term spatial semantics. The original backbone and neck can be represented formally as,

$$\begin{aligned}&\mathcal{G}_n(\mathcal{G}_b(\mathcal{S}, W_b), W_n)\\&= \mathcal{G}_{n+b}(\mathcal{S}, W_{n+b})\\&= \mathcal{G}_{fuse}(\mathcal{G}^{short}_{n+b}(I_t), \mathcal{G}^{long}_{n+b}(I_{t-\delta t}, \ldots, I_{t-N\delta t})),\end{aligned}$$

where $\mathcal{G}_{fuse}$ represents the LSFM-Lf-Dil of LongShortNet. $\mathcal{G}^{short}_{n+b}$ and $\mathcal{G}^{long}_{n+b}$ denote the ShortPath and LongPath of LongShortNet, which are used for feature extraction of the current and historical feature, respectively. Note that their weights are shared.

Finally, the dual-path network is formulated as,

$$\begin{aligned}\mathcal{T} &= \mathcal{F}(\mathcal{S}, W)\\&= \mathcal{G}_h(\mathcal{G}_n(\mathcal{G}_b(\mathcal{S}, W_b), W_n), W_h)\\&= \mathcal{G}_h(\mathcal{G}_{fuse}(\mathcal{G}^{short}_{n+b}(I_t), \mathcal{G}^{long}_{n+b}(I_{t-\delta t}, \ldots, I_{t-N\delta t}))),\end{aligned}$$

where the proposed dual-path architecture effectively addresses complex motion scenarios and offers a sophisticated solution for object detection in video sequences.

3.2 Asymmetric Knowledge Distillation

The ability to retain long-term spatiotemporal knowledge through fused features lends strength to forecasting, yet achieving streaming perception remains a daunting task. Drawing inspiration from knowledge distillation, we've fashioned an asymmetric distillation strategy, transferring "future knowledge" to the present frame. This assists the model in honing its accuracy in streaming perception without the burden of additional inference costs.

Given the asymmetric input nature of the teacher and student networks, a sizable gap emerges in their feature distributions, thus impairing the effectiveness of distillation at the

feature level. Logits-based distillation primarily garners performance improvements by harmonizing the teacher model's response-based knowledge, which aligns knowledge distribution at the semantic level. This simplifies the optimization process for asymmetric distillation. As a result, we've engineered a distillation module to convey rich semantic and localization knowledge from the teacher (the future) to the student (the present).

The asymmetric distillation is depicted in Fig. 3. The teacher model is a still image detector that takes I_{t+1} as input and produces logits for I_{t+1}. The student model is a standard streaming perception pipeline that uses historical frames I_{t-1}, ..., I_{t-N} and the current frame I_t as input to forecast the results of the arriving frame I_{t+1}. The logits produced by the teacher and student are represented by $\mathcal{T}_{stu} = \{F^{cls}_{stu}, F^{reg}_{stu}, F^{obj}_{stu}\}$, and $\mathcal{T}_{tea} = \{F^{cls}_{tea}, F^{reg}_{tea}, F^{obj}_{tea}\}$, where $F^{cls}_{\cdot}$, $F^{reg}_{\cdot}$, and $F^{obj}_{\cdot}$ correspond to the classification, objectness, and regression logits features, respectively. The Asymmetric Knowledge Distillation, AKDM$(\cdot)$, is mathematically formulated as,

$$\begin{aligned}&\text{AKDM}(\mathcal{T}_{stu}, \mathcal{T}_{tea})\\ &= \mathcal{L}_{cls}(F^{cls}_{stu}, F^{cls}_{tea}) + \mathcal{L}_{obj}(F^{obj}_{stu}, F^{obj}_{tea}) + \mathcal{L}_{reg}(\hat{F}^{reg}_{stu}, \hat{F}^{reg}_{tea}),\end{aligned}$$

where $\mathcal{L}_{cls}(\cdot)$ and $\mathcal{L}_{obj}(\cdot)$ are Mean Square Error (MSE) loss functions, and $\mathcal{L}_{reg}(\cdot)$ is the GIoU loss [Rezatofighi *et al.*, 2019]. $\hat{F}^{reg}_{stu}$ and $\hat{F}^{reg}_{tea}$ represent the positive samples of the regression logit features, filtered using the OTA assignment method as in YOLOX [Ge *et al.*, 2021b]. It is worth noting that location knowledge distillation is only performed on positive samples to avoid noise from negative ones.

3.3 K-step Streaming Metric

The Streaming Average Precision (sAP) metric is a prevalent tool used to gauge the precision of Streaming Perception systems [Li *et al.*, 2020]. This metric gauges precision by juxtaposing real-world ground truth with system-generated results, factoring in process latency.

Two primary methodologies exist in this domain: non-real-time and real-time. For non-real-time methods, as depicted in Fig.5(a), the sAP metric calculates precision by comparing the current frame I_t results with the ground truth of the following frame I_{t+2}, post processing of frame I_t. Conversely, real-time methods, as demonstrated in Fig. 5(b), conclude the processing of the current frame I_t prior to the next frame I_{t+1} arrival. Our proposed method, DAMO-StreamNet, is a real-time method, adhering to the pipeline outlined in Fig. 5(b).

Though the sAP metric effectively evaluates the short-term forecasting capability of algorithms, it falls short in assessing their long-term forecasting prowess— a critical factor in real-world autonomous driving scenarios. In response, we introduce the K-step Streaming metric, an expansion of the sAP metric, specifically tailored to evaluate long-term performance. As depicted in Fig. 5(c), the algorithm projects the results of the upcoming two frames, and the cycle continues. The projection of the next K frames is represented as "K-sAP", as shown in Fig. 5(d). Consequently, the standard sAP metric translates to 1-sAP in the K-step metric context.

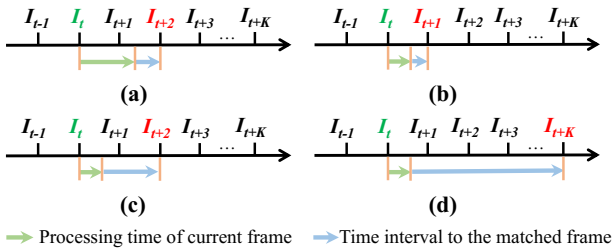


Figure 5: Illustration of matching rules under different metrics. The frames in green font denote the current frame and the frames in red font denote the frames matched with the current frame under the specific metric. (a) Matching result of non-real-time methods under 1-sAP. (b) Matching result of real-time methods under 1-sAP. (c) Matching result of real-time methods under 2-sAP. (d) Matching result of real-time methods under K-sAP.

4 Experiments

4.1 Dataset and Metric

Dataset. We utilized the Argoverse-HD dataset, which comprises various urban outdoor scenes from two US cities. The dataset contains detection annotations and center RGB camera images, which were used in our experiments. We adhered to the train/validation split proposed by Li et al. [Li *et al.*, 2020], with the validation set consisting of 15k frames.

Evaluation Metrics. We employed the streaming Average Precision (sAP) metric to evaluate performance. The sAP metric calculates the average mAP over Intersection over Union (IoU) thresholds ranging from 0.5 to 0.95, as well as APs, APm, and APl for small, medium, and large objects, respectively. This metric has been widely used in object detection, including in previous works such as [Li *et al.*, 2020; Yang *et al.*, 2022a].

4.2 Implementation Details

We pretrained the base detector of our DAMO-StreamNet on the COCO dataset [Lin *et al.*, 2014], following the methodology of StreamYOLO [Yang *et al.*, 2022a]. We then trained DAMO-StreamNet on the Argoverse-HD dataset for 8 epochs with a batch size of 32, using 4 V100 GPUs. For convenient comparison with recent state-of-the-art models [Yang *et al.*, 2022a; Li *et al.*, 2022], we designed small, medium, and large networks (i.e., DAMO-StreamNet-S, DAMO-StreamNet-M, and DAMO-StreamNet-L). The normal input resolution (600, 960) was utilized unless specified otherwise. We maintained consistency with other hyperparameters from previous works [Yang *et al.*, 2022a; Li *et al.*, 2022]. AK-Distillation is an auxiliary loss for DAMO-StreamNet training, with the weight of the loss set to 0.2/0.2/0.1 for DAMO-StreamNet-S/M/L, respectively.

4.3 Comparison with State-of-the-art Methods

We compared our proposed approach with state-of-the-art methods to evaluate its performance. In this subsection, we directly copied the reported performance from their original papers as their results. The performance comparison was conducted on the Argoverse-HD dataset [Li *et al.*, 2020]. An

Methods	sAP	sAP_{50}	sAP_{75}	sAP_s	sAP_m	sAP_l
Non-real-time detector-based methods						
Streamer (S=900) [Li *et al.*, 2020]	18.2	35.3	16.8	**4.7**	14.4	34.6
Streamer (S=600) [Li *et al.*, 2020]	20.4	35.6	20.8	3.6	18.0	47.2
Streamer + AdaScale [Chin *et al.*, 2019; Ghosh *et al.*, 2021]	13.8	23.4	14.2	0.2	9.0	39.9
Adaptive Streamer [Ghosh *et al.*, 2021]	**21.3**	**37.3**	**21.1**	4.4	**18.7**	**47.1**
Real-time detector-based methods						
StreamYOLO-S [Yang *et al.*, 2022a]	28.8	50.3	27.6	9.7	30.7	53.1
StreamYOLO-M [Yang *et al.*, 2022a]	32.9	54.0	32.5	12.4	34.8	58.1
StreamYOLO-L [Yang *et al.*, 2022a]	36.1	57.6	35.6	13.8	37.1	63.3
LongShortNet-S [Li *et al.*, 2022]	29.8	50.4	29.5	11.0	30.6	52.8
LongShortNet-M [Li *et al.*, 2022]	34.1	54.8	34.6	13.3	35.3	58.1
LongShortNet-L [Li *et al.*, 2022]	37.1	57.8	37.7	15.2	37.3	63.8
DAMO-StreamNetNet-S (Ours)	31.8	52.3	31.0	11.4	32.9	58.7
DAMO-StreamNetNet-M (Ours)	35.7	56.7	35.9	14.5	36.3	63.3
DAMO-StreamNetNet-L (Ours)	**37.8**	**59.1**	**38.6**	**16.1**	**39.0**	**64.6**
Large resolution						
StreamYOLO-L ‡	41.6	65.2	43.8	23.1	44.7	60.5
LongShortNet-L †	42.7 (+1.1)	65.4 (+0.2)	**45.0** (+1.2)	23.9 (+0.8)	44.8 (+0.1)	61.7 (+1.2)
DAMO-StreamNet-L † (Ours)	**43.3** (+1.7)	**66.1** (+0.9)	44.6 (+0.8)	**24.2** (+1.1)	**47.3** (+2.6)	**64.1** (+3.6)

Table 1: Comparison with both non-real-time and real-time state-of-the-art (SOTA) methods on the Argoverse-HD benchmark dataset. The symbol '‡' denotes the use of a large size (1200, 1920) and extra data. The symbol '†' denotes the use of a large size (1200, 1920) without the use of extra data. The best results for each setting are shown in green. The largest increments of the large resolution setting are shown in red.

overview of the results reveals that our proposed DAMO-StreamNet with an input resolution of 600 $\times$ 960 achieves 37.8% sAP, outperforming the current state-of-the-art methods by a significant margin. For the large-resolution input of 1200 $\times$ 1920, our DAMO-StreamNet attains 43.3% sAP without extra training data, surpassing the state-of-the-art work StreamYOLO, which was trained with large-scale auxiliary datasets. This clearly demonstrates the effectiveness of the systematic improvements in DAMO-StreamNet.

Compared to StreamYOLO and LongShortNet, DAMO-StreamNet-L achieves absolute improvements of 3.6% and 2.4% under the sAPL metric, respectively. This also provides substantial evidence that the features produced by DRFPN offer a self-adaptive and sufficient size of the receptive field for large-sized objects. It is worth noting that DAMO-StreamNet experiences a slight decline compared to LongShortNet under the stricter metric sAP75. This observation suggests that although the dynamic receptive field achieves a sufficient receptive field for different scales of objects, it is not as accurate as fixed kernel-size ConvNets. The offset prediction in the deformable convolution layer may not be precise enough for high-precision scenarios. In other words, better performance could be achieved if this issue is addressed, and we leave this for future work.

4.4 Ablation Study

Investigation of DRFPN. To verify the effectiveness of DRFPN, we use StreamYOLO [Yang *et al.*, 2022a] and LongShortNet [Li *et al.*, 2022] as baselines and integrate them with the proposed DRFPN, respectively. The experimental results are listed in Table 2. It is evident that DRFPN significantly improves the feature aggregation capability of the baselines. Particularly, the small-scale baseline models equipped with DRFPN achieve improvements of 1.9% and 1.7%, separately. This also demonstrates that the dynamic receptive field is crucial for the stream perception task. More importantly, DRFPN enhances the performance of LongShortNet, which suggests that the temporal feature alignment capacity is also augmented by the dynamic receptive field mechanism.

Methods	S	M	L
Equip StreamYOLO with our DRFPN			
StreamYOLO	28.7	33.5	36.1
+DRFPN	**30.6** (+1.9)	**35.1** (+1.6)	**36.7** (+0.6)
LongShortNet Equipped with our DRFPN			
LongShortNet	29.8	34.0	36.7
+DRFPN	**31.5** (+1.7)	**35.7** (+1.7)	**37.5** (+0.8)

Table 2: Ablation study of the base detector on the Argoverse-HD dataset. The best results for each subset and the corresponding increments are shown in green font and red font, respectively.

Investigation of Temporal Range. To isolate the influence of temporal range, we conduct an ablation study on N and δt, as listed in Table 3. (0, -) represents the model utilizing only the current frame as input. It is evident that increasing the number of input frames can enhance the model's performance, with the best results obtained when N is equal to 2, 2, and 3 for DAMO-StreamNet-S/M/L, respectively. However, as the number of input frames continues to increase, the performance experiences significant declines. Intuitively, longer temporal information should be more conducive to forecasting, but the effective utilization of long-term temporal information remains a critical challenge worth investigating.

Investigation of AK-Distillation. AK-Distillation is a cost-free approach for enhancing the streaming perception pipeline, and we examine its impact. We perform AK-Distillation with various lengths of temporal modeling and scales of DAMO-StreamNet. As the results listed in Ta-

$(N, \delta t)$	StreamNet-S	StreamNet-M	StreamNet-L
(0, -)	28.1	32.0	34.2
(1, 1)	30.6	35.1	36.7
(1, 2)	31.2	34.5	37.1
(2, 1)	31.2	**35.7** (+3.7)	**37.5** (+3.3)
(2, 2)	31.4 (+3.3)	35.4 (+3.4)	37.2
(3, 1)	**31.5** (+3.4)	35.3	37.2
(3, 2)	31.2	35.1	37.4 (+3.2)
(4, 1)	31.1	35.0	37.1
(4, 2)	30.7	35.2	36.5
(5, 1)	31.1	35.0	**37.5** (+3.3)
(5, 2)	30.9	34.7	36.9

Table 3: Exploration of N and δt on the Argoverse-HD dataset. StreamNet denotes our DAMO-StreamNet. The best two results and the worst one are shown in green font, blue font, and purple font, respectively. The best increments are shown in red font.

Methods	S	M	L
D-SN (N=1)	30.6	35.1	36.7
D-SN (N=1)+AK-D	31.5 (**+0.9**)	35.3 (**+0.2**)	37.1 (**+0.4**)
D-SN (N=2/3)	31.5	**35.7**	37.5
D-SN (N=2/3)+AK-D	**31.8** (+0.3)	35.5 (-0.2)	**37.8** (+0.3)

Table 4: Ablation study of our proposed models. D-SN and AK-D represent DAMO-StreamNet and AK-Distillation, respectively. The best results and the largest increments are shown in green font and red font, respectively.

ble 4 indicate, AK-Distillation yields improvements of 0.2% to 0.9% for the DAMO-StreamNet configured with $N = 1$ short-term temporal modeling. This demonstrates that AK-Distillation can effectively transfer "future knowledge" from the teacher to the student. For the DAMO-StreamNet with the setting of $N = 3$, AK-Distillation improves DAMO-StreamNet-S/L by only 0.3%, but results in a slight decline for the medium-scale model. The limited improvement for long-term DAMO-StreamNet is due to the narrow performance gap between the teacher and student, and the relatively high precision is difficult to further enhance.

Investigation of K-step Streaming Metric. We evaluate DAMO-StreamNet with settings $N = 1$ and $N = 2/3$ under the new metric sAP_k, where k ranges from 1 to 6. The results are listed in Table 5. It is clear that the performance progressively declines as k increases, which also highlights the challenge of long-term forecasting. Another observation is that the longer time-series information leads to better performance under the new metric.

Inference Efficiency Analysis. Although the proposed DRFPN has a more complex structure compared to PAFPN, DAMO-StreamNet still maintains real-time streaming perception capabilities. For long-term fusion, we adopt the buffer mechanism from StreamYOLO [Yang *et al.*, 2022a], which incurs only minimal additional computational cost for multi-frame feature fusion.

5 Conclusion

Our research presents DAMO-StreamNet, a novel and robust framework integrating cutting-edge technologies from the YOLO series. Key innovations include (1) a robust neck structure using deformable convolution, (2) a dual-branch design for enhanced time-series data analysis, (3) logit-level distillation, and (4) a dynamic real-time prediction mechanism. Comparison with existing methods on the Argoverse-HD dataset clearly shows DAMO-StreamNet's superiority.

K-Step Metric		StreamNet (N=1)	StreamNet (N=2/3)
S	sAP_1	30.6	**31.5** (+0.9)
	sAP_2	28.3	29.8 (**+1.5**)
	sAP_3	24.9	25.9 (+1.0)
	sAP_4	22.1	23.3 (+1.2)
	sAP_5	21.0	21.8 (+0.8)
	sAP_6	18.8	20.0 (+1.2)
M	sAP_1	35.1	**35.7** (+0.6)
	sAP_2	31.9	32.8 (**+0.9**)
	sAP_3	28.8	29.2 (+0.4)
	sAP_4	25.7	25.9 (+0.2)
	sAP_5	23.2	23.4 (+0.2)
	sAP_6	21.5	22.0 (+0.5)
L	sAP_1	36.7	**37.5** (**+0.8**)
	sAP_2	33.2	33.9 (+0.7)
	sAP_3	29.8	30.6 (**+0.8**)
	sAP_4	27.1	27.2 (+0.1)
	sAP_5	24.2	25.0 (**+0.8**)
	sAP_6	22.3	22.7 (+0.4)

Table 5: Exploration study of K-sAP on the Argoverse-HD dataset. Here, our proposed model DAMO-StreamNet is denoted as StreamNet. The best results and largest increments for each subset are shown in green and red font, respectively.

Methods	S	M	L
LongShortNet (N=1)	14.2	17.3	19.7
LongShortNet (N=3)	14.6	17.5	19.8
DAMO-StreamNet (N=1)	21.0	24.2	26.2
DAMO-StreamNet (N=3)	21.3	24.3	26.6

Table 6: Ablation study of inference time (ms) on V100.

Acknowledgments

Zhi-Qi Cheng's research in this project was supported by the US Department of Transportation, Office of the Assistant Secretary for Research and Technology, under the University Transportation Center Program (Federal Grant Number 69A3551747111), as well as Intel and IBM Fellowships.

Contribution Statement

Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, and Wangmeng Xiang contributed equally as co-first authors to this work, in random order. They were involved in the study design, experiments, manuscript writing, and discussions. The manuscript underwent review by Binghui Chen, Bin Luo, Yifeng Geng, and Xuansong Xie. Zhi-Qi Cheng, as the corresponding author, managed the entire project. Parts of the work were completed through Jun-Yan He's remote collaboration with Carnegie Mellon University.

References

[Belhassen *et al.*, 2019] Hatem Belhassen, Heng Zhang, Virginie Fresse, and El-Bay Bourennane. Improving video object detection by seq-bbox matching. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pages 226–233, 2019.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[Chen *et al.*, 2017] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

[Chen *et al.*, 2018] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7814–7823, 2018.

[Chen *et al.*, 2021] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7028–7036, 2021.

[Chen *et al.*, 2023] Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, Wei Liu, Zhi-Qi Cheng, Hanbing Liu, Bin Luo, Yifeng Geng, and Xuansong Xie. Hdformer: High-order directed transformer for 3d human pose estimation. *arXiv preprint arXiv:2302.01825*, 2023.

[Cheng *et al.*, 2017a] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4048–4056, 2017.

[Cheng *et al.*, 2017b] Zhi-Qi Cheng, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. On the selection of anchors and targets for video hyperlinking. In *Proceedings of the 2017 acm on international conference on multimedia retrieval (ICMR)*, pages 287–293, 2017.

[Cheng *et al.*, 2018] Zhi-Qi Cheng, Xiao Wu, Siyu Huang, Jun-Xiu Li, Alexander G Hauptmann, and Qiang Peng. Learning to transfer: Generalizable attribute learning with multitask neural model search. In *Proceedings of the ACM international conference on Multimedia (ACM MM)*, pages 90–98, 2018.

[Cheng *et al.*, 2019a] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G Hauptmann. Learning spatial awareness to improve crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6152–6161, 2019.

[Cheng *et al.*, 2019b] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, Jun-Yan He, and Alexander G Hauptmann. Improving the learning of multi-column convolutional neural network for crowd counting. In *Proceedings of the ACM international conference on multimedia (ACM MM)*, pages 1897–1906, 2019.

[Cheng *et al.*, 2022] Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 3272–3281, 2022.

[Chin *et al.*, 2019] Ting-Wu Chin, Ruizhou Ding, and Diana Marculescu. Adascale: Towards real-time video object detection using adaptive scaling. In *Proceedings of Machine Learning and Systems (MLSys)*, 2019.

[Ding *et al.*, 2019] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In *International Conference on Computer Vision (ICCV)*, pages 1911–1920, 2019.

[Ding *et al.*, 2021a] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10886–10895, 2021.

[Ding *et al.*, 2021b] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13733–13742, 2021.

[Gao *et al.*, 2021] Kaifeng Gao, Long Chen, Yifeng Huang, and Jun Xiao. Video relation detection via tracklet based visual transformer. In *Proceedings of ACM Conference on Multimedia (ACM MM)*, pages 4833–4837. ACM, 2021.

[Ge *et al.*, 2021a] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. OTA: optimal transport assignment for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 303–312, 2021.

[Ge *et al.*, 2021b] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *CoRR*, abs/2107.08430, 2021.

[Ghiasi *et al.*, 2019] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. NAS-FPN: learning scalable feature pyramid architecture for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7036–7045, 2019.

[Ghosh *et al.*, 2021] A. Ghosh, A. Nambi, A. Singh, and et al. Adaptive streaming perception using deep reinforcement learning. *CoRR*, abs/2106.05665, 2021.

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

[He *et al.*, 2021] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. Db-lstm: Densely-connected bi-directional lstm for human action recognition. *Neurocomputing*, 444:319–331, 2021.

[Heo *et al.*, 2019] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3779–3787, 2019.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[Huang *et al.*, 2018] Siyu Huang, Xi Li, Zhi-Qi Cheng, Zhongfei Zhang, and Alexander Hauptmann. Gnas: A greedy neural architecture search method for multi-attribute learning. In *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*, pages 2049–2057, 2018.

[Huang *et al.*, 2022] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Mingqian Tang, Ziwei Liu, and Marcelo H. Ang Jr. Tada! temporally-adaptive convolutions for video understanding. In *Proceedings of International Conference on Learning Representations, (ICLR)*, 2022.

[Jiang *et al.*, 2022] Yiqi Jiang, Zhiyu Tan, Junyan Wang, Xiuyu Sun, Ming Lin, and Hao Li. Giraffedet: A heavy-neck paradigm for object detection. In *International Conference on Learning Representations (ICLR)*, 2022.

[Kang *et al.*, 2016] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 817–825, 2016.

[Kim and Lee, 2020] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision(ECCV)*, volume 12370, pages 355–371, 2020.

[Lan *et al.*, 2022] Jin-Peng Lan, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Xu Bao, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Procontext: Exploring progressive context transformer for tracking. *arXiv preprint arXiv:2210.15511*, 2022.

[Li *et al.*, 2020] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12347, pages 473–488, 2020.

[Li *et al.*, 2022] Chenyang Li, Zhi-Qi Cheng, Jun-Yan He, Pengyu Li, Bin Luo, Han-Yuan Chen, Yifeng Geng, Jin-Peng Lan, and Xuansong Xie. Longshortnet: Exploring temporal and semantic features fusion in streaming perception. *arXiv preprint arXiv:2210.15518*, 2022.

[Lin *et al.*, 2014] T. Lin, M. Maire, S. Belongie, and et al. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944. IEEE Computer Society, 2017.

[Lin *et al.*, 2020] Lijian Lin, Haosheng Chen, Honglun Zhang, Jun Liang, Yu Li, Ying Shan, and Hanzi Wang. Dual semantic fusion network for video object detection. In *ACM International Conference on Multimedia (ACM MM)*, pages 1855–1863, 2020.

[Liu *et al.*, 2018] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768. Computer Vision Foundation / IEEE Computer Society, 2018.

[Liu *et al.*, 2020] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2020.

[Mirzadeh *et al.*, 2020] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pages 5191–5198, 2020.

[Rezatofighi *et al.*, 2019] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019.

[Sun *et al.*, 2021] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. MAMBA: multi-level aggregation via memory bank for video object detection. In *Proceedings of AAAI Conference on Artificial Intelligence, (AAAI)*, pages 2620–2627, 2021.

[Tan *et al.*, 2020] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, 2020.

[Tu *et al.*, 2023] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. *arXiv preprint arXiv:2304.10465*, 2023.

[Vasu *et al.*, 2022] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. An improved one millisecond mobile backbone. *CoRR*, abs/2206.04040, 2022.

[Wang *et al.*, 2018] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11217, pages 557–573, 2018.

[Wang *et al.*, 2021] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 43(10):3349–3364, 2021.

[Wang *et al.*, 2022] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *CoRR*, abs/2207.02696, 2022.

[Yang *et al.*, 2022a] Jinrong Yang, Songtao Liu, Zeming Li, Xiaoping Li, and Jian Sun. Real-time object detection for streaming perception. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5395, 2022.

[Yang *et al.*, 2022b] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4643–4652, 2022.

[Yao *et al.*, 2021] Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3591–3600, 2021.

[Zhou *et al.*, 2022] Yuxuan Zhou, Chao Li, Zhi-Qi Cheng, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.

[Zhu *et al.*, 2017a] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of IEEE Conference on Computer Vision (ICCV)*, pages 408–417, 2017.

[Zhu *et al.*, 2017b] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4141–4150, 2017.