When Pedestrian Detection Meets Multi-Modal Learning: Generalist Model and Benchmark Dataset

Yi Zhang[2], Wang Zeng[2], Sheng Jin[3,2], Chen Qian[1,2] ✉
Ping Luo[3,4], and Wentao Liu[2]

[1] Tsinghua University [2] SenseTime Research and Tetras.AI
[3] The University of Hong Kong [4] Shanghai AI Laboratory
yizhang.bham.uk@outlook.com, qianc18@mails.tsinghua.edu.cn

Abstract. Recent years have witnessed increasing research attention towards pedestrian detection by taking the advantages of different sensor modalities (*e.g.* RGB, IR, Depth, LiDAR and Event). However, designing a unified generalist model that can effectively process diverse sensor modalities remains a challenge. This paper introduces MMPedestron, a novel generalist model for multimodal perception. Unlike previous specialist models that only process one or a pair of specific modality inputs, MMPedestron is able to process multiple modal inputs and their dynamic combinations. The proposed approach comprises a unified encoder for modal representation and fusion and a general head for pedestrian detection. We introduce two extra learnable tokens, *i.e.* MAA and MAF, for adaptive multi-modal feature fusion. In addition, we construct the MMPD dataset, the first large-scale benchmark for multi-modal pedestrian detection. This benchmark incorporates existing public datasets and a newly collected dataset called EventPed, covering a wide range of sensor modalities including RGB, IR, Depth, LiDAR, and Event data. With multi-modal joint training, our model achieves state-of-the-art performance on a wide range of pedestrian detection benchmarks, surpassing leading models tailored for specific sensor modality. For example, it achieves 71.1 AP on COCO-Persons and 72.6 AP on LLVIP. Notably, our model achieves comparable performance to the InternImage-H model on CrowdHuman with 30× smaller parameters. Codes and data are available at `https://github.com/BubblyYi/MMPedestron`.

Keywords: Pedestrian Detection · Multi-Modal Learning

1 Introduction

Pedestrian detection [13] has long been a hot research topic in computer vision due to its various applications, including autonomous driving, robotics and video surveillance. Traditional pedestrian detection mostly focus on single-modal

✉ : Corresponding author.

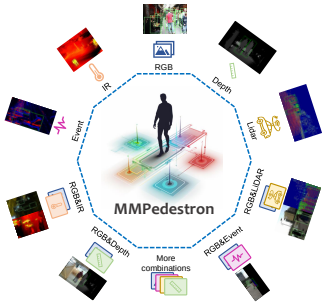

Fig. 1: MMPedestron unifies diverse modality inputs, including RGB, IR, Event, Depth and LiDAR, for pedestrian detection.

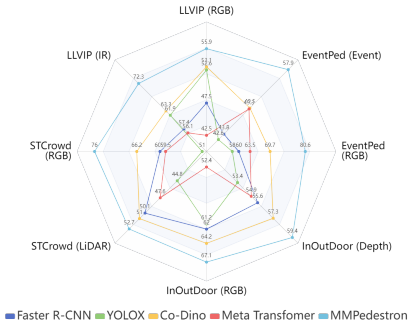

Fig. 2: Performance on diverse datasets and modalities. MMPedestron outperforms leading models trained on the specific dataset and modality.

RGB images as the input. However, RGB based pedestrian detection methods face great challenges in complex scenarios (*e.g.* background clutter and adverse lighting conditions). With the rapid development and application of sensing hardware, multi-modal learning has attracted increasing research attention. Different types of sensors can supply RGB images with rich complementary information and bring remarkable benefits for pedestrian detection. For example, Infrared Radiation (IR) sensor detects heat radiation of pedestrians which is helpful for detection in a dark environment. Time of Flight (ToF) and LiDAR sensors provide additional depth information of the scene.

Previous studies mainly focus on designing specific models for one single or a pair of modality inputs. The development of a unified model that can effectively incorporate various sensor modalities in multi-modal pedestrian detection poses several challenges. Firstly, existing benchmarks for pedestrian detection primarily focus on a single or a pair of sensor modalities, lacking a comprehensive benchmark that can fairly and comprehensively evaluate various methods across diverse application scenarios. Secondly, previous multi-modal fusion methods are often tailored for specific modality pairs (*e.g.* RGB-D or RGB-T), yet hard to be extended to operate with other modality combinations. For example, those models trained for RGB-D data are not applicable to the RGB-T data. Consequently, multiple models are required to deal with different modality combinations, resulting in unnecessary system complexity and inefficiency. In addition, previous fusion methods assume the availability of all modalities and do not account for scenarios where certain modalities may be missing, exacerbating the problem. Lastly, different modality-specific pedestrian datasets are collected from various domains and designed for specific application scenarios (*e.g.* LLVIP [20] for surveillance viewpoints, Waymo [43] for automobiles, InOutDoor [34] for robotics). As a result, previous pedestrian detectors trained on one specific modality lack generalization capabilities across different domains.

In this paper, we make contributions to the field of multi-modal pedestrian detection by introducing both a benchmark dataset and a generalist model. Firstly, we address the lack of a comprehensive benchmark for multi-modal pedestrian detection by constructing the MMPD benchmark. This benchmark dataset is derived from existing public datasets [10, 20, 30, 34, 42]. To address the lack of paired RGB-Event data in the community and to enhance the MMPD benchmark's diversity, we also introduce a new RGB-Event pedestrian detection dataset called EventPed. Our MMPD dataset is diverse in two aspects. **(1) Modality.** MMPD dataset covers multiple sensor modalities, such as RGB, IR, Depth, LiDAR, and Event data, and diverse modality combinations, including RGB+IR, RGB+Depth, RGB+LiDAR, and RGB+Event. **(2) Scenario.** Unlike previous datasets collected under a specific scenario, MMPD encompasses various scenarios, including surveillance, automobile, robotics, outdoor and indoor environments. The diversity of modality and scenario make it possible to develop and evaluate the generalist multi-modal pedestrian detection model.

We propose MMPedestron, a generalist multi-modal pedestrian detection model designed to handle diverse input modalities and scenarios. MMPedestron consists of a unified multi-modal encoder and a detection head. The unified encoder transfer multi-modal input to vision tokens, which are combined with a Modality Aware Fuser (MAF) and a Modality Aware Abstractor (MAA) token to form a hybrid token sequence. The hybrid token sequence is processed by transformer blocks and transferred to unified tokens by modality unifier module. The unified tokens are then passed to the detection head for final predictions. By training on data with different modalities, MMPedestron achieves state-of-the-art performance on various pedestrian detection benchmarks, surpassing models tailored for specific sensor modalities. Furthermore, we highlight several noteworthy properties of MMPedestron: (1) **Flexibility:** MMPedestron exhibits the capability to handle diverse input modalities and their dynamic combinations, allowing for versatility in different applications. (2) **Scalability:** With efficient weight sharing, MMPedestron can seamlessly accommodate an increasing number of modalities without a proportional growth in parameters, demonstrating excellent scalability. (3) **Generalization ability:** The diversity of the MMPD dataset enables MMPedestron to exhibit strong generalization abilities across various domains and scenarios.

Our main contributions can be summarized as follows:

- The introduction of MMPD dataset, a large-scale multi-modal pedestrian detection benchmark, which serves as a standardized evaluation platform for multi-modal pedestrian detection methods.
- Pioneering the concept of generalist multi-modality pedestrian detection through the development of the MMPedestron model. This model is designed to handle diverse input modalities and scenarios, showcasing remarkable flexibility, scalability, and generalization ability.
- Experimental results showcase that our model achieves state-of-the-art performance across a wide range of pedestrian detection benchmarks, outperforming current leading models tailored for specific sensor modality.

2 Related Works

2.1 Multi-Modal Object Detection

While RGB images provide substantial texture information and details for pedestrian detection, the inclusion of multi-modal data is desirable to achieve more reliable results in challenging conditions, such as extreme lighting, occlusions, and fast motion. Existing methods have developed various strategies to fuse information from multiple modalities, namely early-fusion, late-fusion, and mid-fusion. **Early-fusion** (pixel fusion) concatenates data from different modalities and process it with regular object detectors [29, 38, 39]. **Late-fusion** (decision fusion) feeds the inputs of two modalities separately into two unimodal object detection models, and then fusing predicted bounding boxes using statistical methods [6, 25, 45]. Early-fusion and late-fusion approaches are straightforward, however they ignore the correlations between modalities. **Mid-fusion** (feature fusion) fuses the features extracted from multiple modalities and predicts bounding boxes from the fused feature. Most research on multi-modal object detection focuses on mid-fusion [3, 31, 36, 37, 47, 55], as this strategy enables the deep exploration of the correlations between modalities.

Our MMPedestron model falls within the mid-fusion category. Previous mid-fusion methods usually use separated branches for different modalities. However, we use a unified encoder for all modalities, demonstrating better scalability. In addition, previous fusion approaches primarily focus on bi-modal features and do not consider diverse modal combinations. In contrast, our model offers flexibility in handling diverse combinations of modalities and scenarios.

2.2 Multi-Modal Benchmarks

To facilitate the development of generalist multi-modality pedestrian detection models, it is crucial to have diverse data comprising various modalities. As depicted in Table 1, while there exists large repositories of annotated RGB-based datasets, there are much fewer annotated data of other modalities (*e.g.* Depth), and even scarcer annotations of modality combinations (*e.g.* RGB + Depth). More importantly, existing multi-modal benchmarks for pedestrian detection typically consist of only a single pair of modalities, *e.g.* LLVIP [20] and InOut-Door [34]. In contrast, MMPD dataset which integrates multiple public datasets, encompasses five modalities and four distinct modal combinations.

Event-based datasets for pedestrian detection. The advantages in handling challenging lighting conditions, high motion, and low latency make event data well-suited for pedestrian detection [11, 17, 35]. However, due to the challenges in data collection and annotation, the availability of annotated event data is significantly limited compared to other common modalities. GEN1 dataset [11] and PEDRo dataset [1] offer manual annotations, however they have relatively low image resolution and lack paired RGB images. DSEC [17] offers paired RGB and event data without object bounding box annotations. A recent work [47] proposes an automated labeling protocol to generate box annotations for DSEC, but

Table 1: Overview of representative multi-modal pedestrian detection datasets. "# Img" means the number of total images. "Label" includes manual labels or pseudo-labels generated by models. MMPD has paired modality data covering all modalities.

Dataset	# Img	Label	RGB	IR	LiDAR	Depth	Event
Caltech [12]	250K	Manual	✓				
CityPersons [58]	5K	Manual	✓				
CrowdHuman [42]	24K	Manual	✓				
Objects365-Persons [41]	133K	Manual	✓				
COCO-Persons [30]	66K	Manual	✓				
LLVIP [20]	15K	Manual	✓	✓			
M3FD [32]	4K	Manual	✓	✓			
FLIR [19]	26K	Manual	✓	✓			
STCrowd [10]	8K	Manual	✓		✓		
Waymo [43]	36K	Manual	✓		✓		
InOutDoor [34]	7K	Manual	✓			✓	
MobilityAids [48]	17K	Manual	✓			✓	
DSEC [17, 47]	11K	Pseudo	✓				✓
PEDRo [1]	27K	Manual					✓
EventPed (Ours)	9K	Manual	✓				✓
MMPD (Ours)	260K	Manual	✓	✓	✓	✓	✓

the utilized data is not publicly released. In contrast, our proposed EventPed dataset overcomes these limitations by providing high-resolution RGB-Event pairs collected in diverse environments, along with comprehensive manual annotations for pedestrian detection.

2.3 Generalist Model

Multi-modal generalist model. The emergence of unified models that incorporate multiple modalities has garnered significant attention due to their exceptional performance across various tasks. ImageBind [18] learns a joint embedding across six different modalities through data pairs of image and other modalities. LanguageBind [61] further improves the joint embedding space by considering language as the bind modality. While these works utilize separate encoders for different modalities, our MMPedestron employs a shared encoder for all modalities. Meta-Transformer [59] leverages a frozen encoder trained on RGB images to perform the perception of multiple modalities. However, Meta-Transformer processes only a single modality per task. In contrast, MMPedestron is capable of handling diverse combinations of multiple modalities. Moreover, Meta-Transformer is exclusively trained on the RGB modality, whereas MMPedestron is trained on a mixture of multiple modalities. **Human-centric generalist model.** Recent studies [8, 21, 23, 46, 52] have explored the development of a generalist model that exploits the commonalities among multiple human-centric tasks. For example, Tang et al. [46] propose HumanBench, a human-centric benchmark comprising six downstream tasks, and train a generalist model based on this benchmark. And UniHCP [8] trains a unified model for human-centric perception using a combination of 33 datasets. While HumanBench and UniHCP primarily focus on the RGB modality to obtain a model suitable for multiple tasks, our focus lies in pedestrian detection and aims to develop a model suitable for multiple modalities.

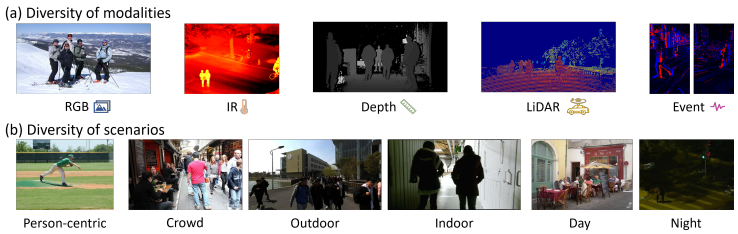

Fig. 3: Overview of our proposed MMPD benchmark. (a) It encompasses a wide range of modalities, such as RGB, IR, Depth, LiDAR, and Event. (b) It includes diverse scenarios, including person-centric v.s. crowd, outdoor v.s. indoor, day v.s. night scenes.

3 MMPD Benchmark

In this paper, we introduce Multi-Modal Pedestrian Detection (MMPD) benchmark based on existing datasets and our newly collected EventPed dataset to comprehensively study the challenging task of multi-modal pedestrian detection. As depicted in Fig. 3, MMPD dataset offers an extensive representation of pedestrians, encompassing various modalities, including RGB, IR, depth, LiDAR, and event, and diverse scenarios, such as different types of occlusion, view-point and illumination conditions.

3.1 Dataset Composition

MMPD is composed of the following datasets: **Objects365-Persons** is derived from Objects365 [41], which is a large-scale RGB-based object detection dataset. We utilize only the training images related to the “person" category. **COCO-Persons** is a subset of COCO [22,30], a well-known RGB-based object detection dataset, which includes images containing the "person" category. **CrowdHuman** [42] is a widely used RGB-based benchmark dataset for pedestrian detection in crowd scenarios. **LLVIP** [20] is a visible-infrared paired dataset for low-light vision, containing $15K$ RGB-IR image pairs. **InOutDoor** [34] is an RGB-Depth paired dataset for pedestrian detection. It contains $6,316$ image pairs for training and $1,028$ image pairs for evaluation. **STCrowd** [10] is an RGB-LiDAR paired dataset for pedestrian perception in crowded scenes. It includes $5,262$ image pairs for training and $2,988$ pairs for evaluation. **EventPed.** To address the lack of paired RGB-Event data, we propose the EventPed dataset for pedestrian detection. More details are presented in Section. 3.2.

3.2 EventPed Dataset

EventPed dataset is a newly collected RGB-event paired dataset focusing on pedestrian detection, which can be useful for robotics, autonomous driving, and surveillance applications. It addresses the scarcity of annotated RGB-event data, facilitating future research and development.

Data collection. The EventPed dataset was collected from March 2023 to July 2023. It encompasses individuals aged between 20 and 70 years recorded in diverse outdoor scenarios such as parks and sidewalks, encompassing both day and night conditions. Prior to recording, informed written consent was obtained from all individuals involved. Our portable data collection device contains a high resolution event camera [15] (IMX636), and a high-quality RGB camera (IMX586). The camera height varied across recordings, and event streams were captured at diverse time intervals ranging from 10 to 20 ms.

Annotation. The dataset underwent manual annotation by well-trained annotators. Each individual in the dataset was exhaustively annotated with a full bounding box. In cases where individuals were partially occluded, annotators were instructed to complete the invisible parts and provide a full bounding box. Quality inspections and manual corrections are performed to ensure the annotation quality. We split the annotated data into a training set with $7,195$ image pairs and a test set with $2,435$ image pairs.

3.3 Evaluation Scenarios

A generalist multi-modal pedestrian detection model is expected to be capable of dealing with both diverse modalities and different modal combinations. So we establish two evaluation scenarios for MMPD dataset. **Unimodal evaluation.** We feed the model with the input in a single modality and evaluate the model on the test set of COCO, CrowdHuman, LLVIP, InOutDoor, STCrowd, and EventPed. For the datasets with multiple modalities, we report the performance with input in both modalities. **Multi-modal evaluation.** We choose four datasets with multiple modalities, *i.e.* LLVIP, InOutDoor, STCrowd, and EventPed, and evaluate the performance of the model with multi-modal input. **Metrics.** Unless otherwise specified, we use COCO AP [30](IoU=0.5:0.95, maxDets=100) as the evaluation metric on all datasets.

4 MMPedestron Model

4.1 Overview

The overview of MMPedestron framework is illustrated in Fig. 4 (a). MMPedestron consists of a unified multi-modal encoder and a detection head. The unified encoder directly takes multi-modal data (*e.g.* RGB and IR data) as the input, and generates unified vision tokens with information aggregated from multiple modalities. These unified tokens are then fed into the detection head to obtain the final result. Our unified tokens are compatible with various pedestrian detection heads, and in our implementation, we choose the recent Co-Dino [63] head for its effectiveness.

4.2 Unified Multi-Modal Encoder

In contrast to previous approaches [2, 56, 57] that employ separate branches for processing multi-modal data, our MMPedestron utilizes a unified transformer

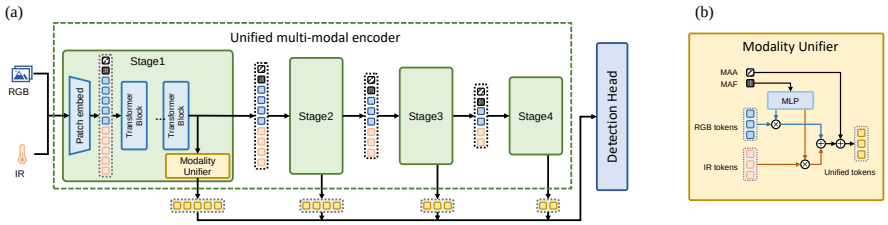

Fig. 4: (a) MMPedestron consists of an unified multi-modal encoder and a detection head. Each stage of the encoder contains a modality-specific patch embedding layer, several transformer blocks and a modality unifier. The resulting unified tokens from multiple stages are fed into the detection head to produce detection results. (b) Modality unifier fuses multi-modal vision tokens with the guidance of MAF and incorporates the domain knowledge of MAA to the output unified tokens. For clarity, we show the case of two modalities.

encoder to handle data from all modalities. As shown in Fig. 4 (a), the unified encoder follows a multi-stage architecture with four hierarchical stages. Each stage contains a modality-specific patch embedding, a series of stacked transformer blocks, and a modality unifier. In line with common practice [14, 33, 50], we convert the input data from each modality into a sequence of vision tokens using a modality-specific patch embedding layer within each stage. Additionally, we prepend two extra learnable tokens, *i.e.* the modality-aware abstractor (MAA) and the modality-aware fuser (MAF), to capture the knowledge of input modality combinations. The multi-modal vision tokens, along with the MAA and MAF tokens, are combined to form a hybrid token sequence, which undergoes further processing by multiple transformer blocks. Our framework is compatible with various commonly used vision transformer blocks, and in this study, we employ the dual vision transformer block [51] for its excellent performance and efficiency. The unified multi-modal encoder offers several advantages over traditional multi-branch architectures. (1) It is more lightweight as different modalities share most of the parameters. (2) It allows the model to learn general knowledge across all modalities, enhancing its ability to generalize and adapt to diverse modalities. (3) It enables more effective and more comprehensive message passing through the attention mechanism within each transformer block.

4.3 Modality Unifier

Given the presence of multi-modal vision tokens, conventional detection heads face difficulties in discerning the optimal utilization of these tokens. To address this issue, we propose a modality unifier module that transforms the hybrid token sequence into a unified token sequence with the same format as standard unimodal vision tokens.

Our unifier module employs two additional learnable tokens to guide the unification process. The **Modality-Aware Fuser (MAF)** token aims to assess the importance or relevance of each modality in the multi-modal fusion pro-

cess. **Modality-Aware Abstractor (MAA)** token aims to collect the domain knowledge related to input modalities.

As illustrated in Fig. 4 (b), when faced with multi-modal inputs, we initially employ a Multilayer Perceptron (MLP) to process the MAF token, obtaining confidence scores for all modalities: $\boldsymbol{c} = Sigmoid(MLP(\boldsymbol{x}_{\text{MAF}}))$, where $\boldsymbol{x}_{\text{MAF}}$ denotes the feature of the MAF token, and $\boldsymbol{c}$ represents the modality confidence, reflecting the importance of each modality. Subsequently, we fuse multi-modal vision tokens through weighted averaging, with weights determined by the predicted modality confidence c. Additionally, we aggregate the information contained in the MAA token by incorporating it into the unified tokens:

$$\mathbf{X}_{\text{uni}} = \frac{\sum_{i=1}^{m} (w_i * c_i * \mathbf{X}_i)}{\sum_{i=1}^{m} (w_i)} + \boldsymbol{x}_{\text{MAA}}, \tag{1}$$

$$\text{where} \quad w_i = \begin{cases} 1, & \mathbf{X}_i \text{ is valid token.} \\ 0, & \mathbf{X}_i \text{ is padded empty token.} \end{cases} \tag{2}$$

$\mathbf{X}_{\text{uni}}$ denotes the features of the unified tokens, and $\boldsymbol{x}_{\text{MAA}}$ is the feature of the MAA token. m is the number of modalities, c_i and X_i denote the confidence and token feature of the ith modality, respectively. w_i is a factor reflecting whether the ith modality is valid. If a modality is missing, we pad the corresponding empty tokens. In such cases, we set w_i to 0, indicating that the modality is absent. Conversely, if a modality is present, we set w_i to 1, denoting its validity. More details are discussed in Section 4.4.

The MAF token and dynamic fusion process in the unifier module allow for adaptive adjustment of the contribution of each modality based on their importance. It enables our model to adaptively allocate attention and resources to different modalities, and to leverage the complementary information provided by each modality. The MAA token provides the domain knowledge related to input modalities. It enhances the model's ability to understand and utilize the specific characteristics of each modality.

4.4 Multi-Modality Training

We utilize a two-stage training scheduler for MMPedestron. In the RGB pretrain stage, we train our MMPedestron on the mixture of large-scale RGB datasets, including Objects365-Persons [41], COCO-Persons [30], and CrowdHuman [42], to learn the general knowledge about human body. In the multi-modal training stage, we train the pretrained model on the mixture of CrowdHuman [42], LLVIP [20], InOutDoor [34], STCrowd [10], and EventPed datasets. The hybrid training data contains diverse modalities and modality combinations, which is essential to a generalist multi-modal model. We treat all modalities as 2D image inputs. For the LiDAR data, we project the 3D points into an image with sparse depth points. For the event data, we integrate the event signals with a time interval to get a 2D image. When dealing with N input modalities, previous approaches typically employ N independent branches. In contrast, the MAA

and MAF enables MMPedestron to handle all modalities with a single shared branch, effectively reducing the parameter number by a factor of N.

We design a modality dropout strategy to enable our MMPedestron model to effectively handle both diverse unimodal inputs and their combinations. This strategy involves randomly dropping a modality from the multi-modal inputs with a probability denoted as p. The model is compelled to process individual modalities as well as their joint representations. When a specific modality is missing in the input, we pad an empty image to the input and mask out all the padded tokens during the process in the encoder, including the transformer blocks and the unifier module. For classification, we use quality focal loss [26] and cross-entropy loss, and for regression, we employ GIOU loss [40] and L_1 loss, following established practices [53, 60, 63].

5 Experiments

We conducted a comprehensive evaluation of our MMPedestron model on multiple challenging datasets, assessing its performance through both unimodal and multi-modal fusion evaluations. Additionally, we examined the transferability of our model using cross-dataset transfer evaluation.

5.1 Implementation Details

MMPedestron adopts Dual ViT [51] as the backbone, pretrained on ImageNet1K. The training process consists of two stages: **RGB Pretrain Stage**. We pretrain MMPedestron on the combined RGB-based dataset for 12 epochs using 64 NVIDIA V100 GPUs. The RGB pretrain stage requires a total of 27,648 GPU hours. **Multi-modal Training Stage**. We train MMPedestron on multi-modality datasets, with modality dropout probability $p = 0.3$. The entire training process comprises 550k iterations using 32 NVIDIA V100 GPUs. The multi-modal training stage requires a total of 1,056 GPU hours. Please refer to Supplementary for more details.

5.2 Unimodal Evaluation

Multi-modal datasets. We first evaluate our MMPedestron model on various multi-modal datasets (*i.e.* LLVIP [20], STCrowd [10], InOutDoor [34], and EventPed) using the unimodal setting. We compare MMPedestron with a range of single-modality detectors, including two-stage detectors *i.e.* Faster R-CNN [39], one-stage detectors *i.e.* YOLOX [16] and query-based detectors *i.e.* Co-Dino [63]. We also compare with a recent multi-modality model Meta-Transformer [59]. As illustrated in Fig. 2, MMPedestron consistently outperforms the competing models across all datasets and modalities. Notably, we evaluate MMPedestron directly on the test set without further dataset-specific fine-tuning. These results demonstrate the general capability of MMPedestron in handling diverse modalities.

Table 2: System-level comparisons with state-of-the-art RGB-based pedestrian detection. † means the models with the composite techniques [27]. ↑ means higher is better, while ↓ means lower is better.

	#Param.	AP ↑
Faster RCNN-80 [39]	42M	51.9
Faster RCNN-Person [39]	41M	54.1
DINO-80 [53]	218M	62.3
DINO-Person [53]	218M	61.8
UniHCP [8]	109M	58.1
InternImage-XL [49]	387M	64.8
Ours (Direct)	62M	68.2
Ours (Finetune)	62M	**71.1**

(a) COCO-Persons `val` dataset

	#Param.	AP ↑	MR^{-2} ↓	JI ↑
DETR [4]	41M	75.9	73.2	74.4
CrowdDet [7]	42M	90.7	41.4	82.3
PEDR [28]	41M	91.6	43.7	83.3
D-DETR [62]	40M	91.5	43.7	83.1
S-RCNN [44]	106M	91.3	44.8	81.3
Iter-D-DETR [60]	207M	94.1	37.7	87.1
PATH [46]	320M	90.8	-	-
UniHCP [8]	109M	92.5	41.6	85.8
InternImage-H [49]	1.09B	95.4	37.9	86.6
InternImage-H† [49]	2.18B	**97.2**	31.1	**89.7**
Ours	62M	97.1	**30.8**	88.0

(b) CrowdHuman dataset

COCO-Persons dataset. To validate the effectiveness of MMPedestron on traditional RGB-based detection, we compare it with state-of-the-art pedestrian detection methods on the widely-used COCO-Persons dataset [30]. We compare MMPedestron against notable models trained on COCO dataset, *i.e.* Faster-RCNN [39] and DINO [53]. Note that these models are originally trained to handle general 80 classes (as indicated by '-80' in Table 2a). To ensure fair comparisons, we utilize MMDetection [5] to re-train these models on "Person" category using the default experimental setting (maked with '-Person' in Table 2a). For fair comparisons under similar model size (number of parameters), we report the results using the ResNet-50 as the backbone. Additionally, we compare MMPedestron with two unified models trained on large-scale datasets: UniHCP [8] and InternImage-XL [49]. UniHCP is pretrained with multi-task learning on a mixture of 33 human-centric datasets, including COCO-Persons dataset. InternImage-XL model is pretrained on ImageNet22k and then fine-tuned on COCO dataset. Note that we report the result of InternImage-XL with Cascade R-CNN [1], because InternImage-H is not publicly available. As shown in Table 2a, our MMPedestron with direct evaluation demonstrates a significant performance margin of 3.4 AP against InternImage-XL, which is 6× larger than MMPedestron. Furthermore, fine-tuning MMPedestron on the COCO-Persons dataset further improves the performance to 71.1 AP. These remarkable results on the COCO-Persons dataset provide strong evidence of the effectiveness of MMPedestron in handling RGB data.

CrowdHuman dataset. In order to assess the efficacy of MMPedestron in handling complex crowd scenarios, we compare our model with the state-of-the-art methods on the CrowdHuman benchmark. The evaluation metrics used in this benchmark include AP, MR^{-2}, and Jaccard index (JI), which are commonly employed in previous studies [7]. As presented in Table 2b, our MMPedestron model outperforms its counterparts by a substantial margin, without increasing

[1] https://github.com/OpenGVLab/InternImage

Table 3: Multi-modality fusion evaluation. AP is reported.

Method	#Param.	LLVIP	STCrowd	InOutDoor	EventPed
Early-Fusion [31]	41M	53.6	54.4	58.3	47.4
FPN-Fusion [47]	65M	57.2	61.5	60.1	61.1
ProbEN [6]	82M	54.8	60.0	62.4	60.1
HRFuser [2]	101M	53.9	49.0	58.6	46.0
CMX [56]	150M	59.6	61.0	62.3	58.0
Ours (RGB Pretrain only)	62M	50.0	59.5	36.8	71.9
Ours (RGB Pretrain + Multi-modal Training)	62M	**72.6**	**74.9**	**65.7**	**79.0**

the model complexity. Specifically, MMPedestron surpasses PATH [46] (ViT-L), UniHCP [8] (ViT-B), and Iter-D-Detr [60] (Swin-L) by 6.3% AP, 4.6% AP and 3.0% AP, respectively. Even when compared to state-of-the-art large-scale model, such as InternImage-H, which is over 30 times larger than our MMPedestron, our model achieves comparable performance. These results validate the ability of MMPedestron to handle challenging crowd scenarios.

5.3 Multi-Modal Evaluation

To assess the capacity to integrate information from multiple modalities, we conducted a comparative analysis between our MMPedestron model and various modality-fusion approaches, including early-fusion [31], mid-fusion [2, 56] and late-fusion approaches [6]. Early-Fusion [31] involves concatenating multi-modal data prior to model input. FPN-Fusion [47] fuses features from various encoder stages through simple addition. And CMX [56] adopts a transformer-based model to incorporate the fusion of RGB with various modalities. It was originally designed for semantic segmentation, but we adapted it to perform pedestrian detection. HRFuser [2] fuses multiple sensors in a multi-resolution fashion with multi-window cross-attention blocks. ProbEN [6] trains separate models for each modality individually and aggregates the predicted bounding boxes of all models with post-processing. As depicted in Table 3, when compared to models separately trained on specific datasets, our MMPedestron model consistently outperforms all fusion methods on diverse datasets without requiring dataset-specific fine-tuning. These results validate the efficacy of MMPedestron in integrating multi-modal information and its ability to generalize across various modality combinations. Moreover, we report the results of our MMPedestron in the RGB pretrain stage, which only uses the RGB modality input. Although RGB pretrain could help learn the general knowledge about human body, multi-modal training significantly improves the detection performance and helps to learn extensive representation of pedestrians, in Table 3. This result also demonstrates the necessity of research on multi-modal pedestrian detection.

5.4 Cross-Dataset Transfer Evaluation

Cross-dataset transfer evaluation aims to measure the model's capability to adapt to new scenarios. We finetune our MMPedestron on the PEDRo [1] dataset and compare it with state-of-the-art methods specifically trained on that dataset.

Table 4: Cross-dataset transfer evaluation.

Method	#Param.	Val AP	Test AP
YOLOv8 [24]	68M	-	58.6
Faster R-CNN [39]	41M	66.8	58.0
YOLOX [16]	99M	73.9	68.8
Co-Dino [63]	64M	73.4	65.4
Meta Transformer [59]	155M	67.5	61.1
Ours (10% train data)	62M	79.3	72.7
Ours	62M	**81.5**	**73.3**

(a) Pedestrian detection on PEDRo dataset (Event).

Method	#Param.	AP50
GAFF [55]	-	72.9
CFT [37]	208M	77.7
CMX [56]	150M	82.2
Ours	62M	**86.4**

(b) Multi-category object detection on FLIR test dataset (RGB + IR).

Table 5: Ablation study on the LLVIP dataset.

MAF	MAA	Fusion	RGB	IR
✗	✗	61.7	47.5	61.2
✓	✗	68.2 (+6.5)	49.8 (+2.3)	68.2 (+7.0)
✓	✓	**68.9** (+0.7)	**52.9** (+3.1)	**68.4** (+0.2)

It is important to note that the PEDRo dataset utilizes a distinct representation for event data, rendering our pre-trained MMPedestron model incompatible for direct evaluation. The results in Table 4a, demonstrate that even with only 10% of the training data, our MMPedestron model achieves state-of-the-art performance on both Val and Test sets. Furthermore, by fine-tuning MMPedestron on the complete training set, we obtain even higher performance, achieving a new state-of-the-art result of 81.5 AP on the Val set and 73.3 AP on the Test set. The quick adaptation to PEDRo dataset showcases the exceptional generalization capacity of our MMPedestron model, which is a result of its multi-modal training on MMPD dataset.

We perform multi-modality fine-tuning experiments on FLIR dataset [19]. Following [37, 56], we conduct experiments on the "aligned" FLIR dataset [54], which consists of 4,192 training pairs and 1,013 testing pairs covering three object categories: 'person', 'bicycle', and 'car'. We extend MMPedestron to support all three categories in FLIR, which further validates its generalization and transfer ability to novel tasks and domains. As shown in the Table 4b, MMPedestron achieves state-of-the-art performance on the FLIR test set.

5.5 Ablation Study

We conduct ablation study on the LLVIP dataset to comprehensively assess the effect of each component in our proposed model in Table 5. The baseline (row #1) utilizes the Dual-ViT backbone with the Co-DINO head. By comparing the performance of various configurations, we observed significant improvements when incorporating Modality-Aware Fusion (MAF). Particularly, the Fusion results showed a notable increase of 6.5 AP, showcasing the effectiveness of MAF in enhancing multi-modality fusion. Additionally, the Modality Attention Abstractor (MAA) component was found to enhance the unification of different modalities. This led to a significant improvement in the lower-performance modality (RGB), with an increase of 3.1 AP. Consequently, the fusion results were further

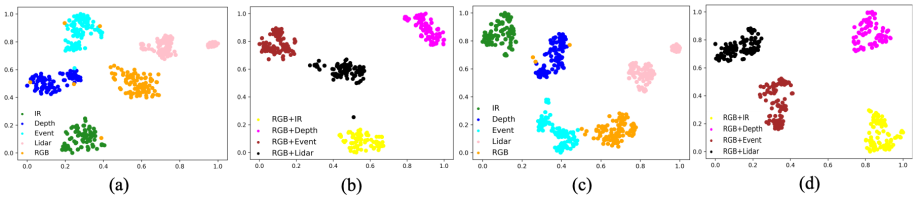

Fig. 5: Visualization of MAF (a, b) and MAA (c, d) tokens. (a,c) are for unimodal inputs, and (b,d) are for multi-modal inputs.

elevated. These findings highlight the importance of both MAF and MAA in our model, as they contribute to improved performance by effectively addressing the challenges associated with multi-modal learning.

5.6 Analysis of MAF and MAA

To gain insights into the properties of our Modality-Aware Fusion (MAF) and Modality-Aware Abstractor (MAA) tokens, we conducted an analysis by visualizing distributions of token features for different input modalities with t-SNE [9]. As shown in Fig. 5 (a) and (c), when processing single-modality inputs, both MAF and MAA samples show clustering patterns corresponding to the respective input modality. This indicates that our MAF and MAA are "modality-aware", and are able to adaptively adjust token features according to the input modality. Similarly, Fig. 5 (b) and (d) demonstrate distinguishable clustered patterns for different modality combinations. This adaptability allows our MMPedestron model to dynamically select appropriate strategies for multi-modal feature fusion according to the specific modality combination. The modality-aware token features ensure the generalization ability of MMPedestron to diverse modalities and modality combinations.

5.7 Runtime Analysis

We conduct the runtime analysis of MMPedstron with a batchsize of 1 on one RTX-4090 GPU in a single thread. MMPedestron achieves near real-time performance (24fps) on the FLIR [19] dataset. Model compilation, pruning, and quantization can further enhance speed, but they are beyond this paper's scope.

6 Conclusion

In this paper, we have presented a pioneering approach to multi-modal pedestrian detection. We introduced the MMPD benchmark, which is the first large-scale benchmark specifically designed for developing and evaluating multi-modal pedestrian detection models. Building upon this benchmark, we proposed MMPedestron, which effectively handles diverse modalities and scenarios. Through comprehensive experiments, we have demonstrated the effectiveness of our MMPedestron model. We hope our work could serve as a valuable foundation for the development of future multi-modal generalist pedestrian detection models.

Acknowledgement. This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622 and 17209324.

References

1. Boretti, C., Bich, P., Pareschi, F., Prono, L., Rovatti, R., Setti, G.: Pedro: an event-based dataset for person detection in robotics. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (Jun 2023)
2. Broedermann, T., Sakaridis, C., Dai, D., Van Gool, L.: Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. IEEE Conf. Intell. Transp. Syst. (2023)
3. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 403–411 (2023)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229. Springer (2020)
5. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Chen, Y.T., Shi, J., Ye, Z., Mertz, C., Ramanan, D., Kong, S.: Multimodal object detection via probabilistic ensembling. In: Eur. Conf. Comput. Vis. pp. 139–158. Springer (2022)
7. Chu, X., Zheng, A., Zhang, X., Sun, J.: Detection in crowded scenes: One proposal, multiple predictions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12214–12223 (2020)
8. Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., Ouyang, W.: Unihcp: A unified model for human-centric perceptions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17840–17852 (2023)
9. Cieslak, M.C., Castelfranco, A.M., Roncalli, V., Lenz, P.H., Hartline, D.K.: t-distributed stochastic neighbor embedding (t-sne): A tool for eco-physiological transcriptomic analysis. Marine genomics **51**, 100723 (2020)
10. Cong, P., Zhu, X., Qiao, F., Ren, Y., Peng, X., Hou, Y., Xu, L., Yang, R., Manocha, D., Ma, Y.: Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19608–19617 (2022)
11. De Tournemire, P., Nitti, D., Perot, E., Migliore, D., Sironi, A.: A large scale event-based detection dataset for automotive. arXiv preprint arXiv:2001.08499 (2020)
12. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 304–311. IEEE (2009)
13. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2011)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Int. Conf. Learn. Represent. (2021)
15. Finateu, T., Niwa, A., Matolin, D., Tsuchimoto, K., Mascheroni, A., Reynaud, E., Mostafalu, P., Brady, F., Chotard, L., LeGoff, F., et al.: 5.10 a 1280x720

back-illuminated stacked temporal contrast event-based vision sensor with 4.86 um pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline. In: IEEE Int. Solid-State Circuits Conf. pp. 112–114 (2020)
16. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
17. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robot. Autom. Lett. **6**(3), 4947–4954 (2021)
18. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15180–15190 (2023)
19. Imaging, T.: Flir data set dataset (2023), `https://universe.roboflow.com/thermal-imaging-0hwfw/flir-data-set`
20. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Int. Conf. Comput. Vis. pp. 3496–3504 (2021)
21. Jin, S., Li, S., Li, T., Liu, W., Qian, C., Luo, P.: You only learn one query: Learning unified human query for single-stage multi-person multi-task human-centric perception. In: Eur. Conf. Comput. Vis. (2024)
22. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Eur. Conf. Comput. Vis. pp. 196–214 (2020)
23. Jin, S., Yao, R., Xu, L., Liu, W., Qian, C., Wu, J., Luo, P.: Unifs: Universal few-shot instance perception with point representations. In: Eur. Conf. Comput. Vis. (2024)
24. Jocher, G., Chaurasia, A., Qiu, J.: Yolo by ultralytics (2023), `https://github.com/ultralytics/ultralytics`
25. Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. Brit. Mach. Vis. Conf. (2018)
26. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Adv. Neural Inform. Process. Syst. **33**, 21002–21012 (2020)
27. Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H.: Cbnet: A composite backbone network architecture for object detection. IEEE Trans. Image Process. **31**, 6893–6906 (2022)
28. Lin, M., Li, C., Bu, X., Sun, M., Lin, C., Yan, J., Ouyang, W., Deng, Z.: Detr for crowd pedestrian detection. arXiv preprint arXiv:2012.06785 (2020)
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2117–2125 (2017)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755. Springer (2014)
31. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. Brit. Mach. Vis. Conf. (2016)
32. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5802–5811 (2022)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Int. Conf. Comput. Vis. pp. 10012–10022 (2021)

34. Mees, O., Eitel, A., Burgard, W.: Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In: IEEE Int. Conf. Intell. Robots Syst. pp. 151–156 (2016)
35. Perot, E., De Tournemire, P., Nitti, D., Masci, J., Sironi, A.: Learning to detect objects with a 1 megapixel event camera. Adv. Neural Inform. Process. Syst. **33**, 16639–16652 (2020)
36. Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv:2111.00273 (2021)
37. Qingyun, F., Zhaokui, W.: Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognition **130**, 108786 (2022)
38. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 779–788 (2016)
39. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inform. Process. Syst. **28** (2015)
40. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 658–666 (2019)
41. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Int. Conf. Comput. Vis. pp. 8430–8439 (2019)
42. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
43. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2446–2454 (2020)
44. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14454–14463 (2021)
45. Takumi, K., Watanabe, K., Ha, Q., Tejero-De-Pablos, A., Ushiku, Y., Harada, T.: Multispectral object detection for autonomous vehicles. In: Proceedings of the on Thematic Workshops of ACM Multimedia. pp. 35–43 (2017)
46. Tang, S., Chen, C., Xie, Q., Chen, M., Wang, Y., Ci, Y., Bai, L., Zhu, F., Yang, H., Yi, L., et al.: Humanbench: Towards general human-centric perception with projector assisted pretraining. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21970–21982 (2023)
47. Tomy, A., Paigwar, A., Mann, K.S., Renzaglia, A., Laugier, C.: Fusing event-based and rgb camera for robust object detection in adverse conditions. In: IEEE Int. Conf. Robot. Autom. pp. 933–939. IEEE (2022)
48. Vasquez, A., Kollmitz, M., Eitel, A., Burgard, W.: Deep detection of people and their mobility aids for a hospital robot. In: European Conference on Mobile Robots (ECMR). pp. 1–7. IEEE (2017)
49. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14408–14419 (2023)

50. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Int. Conf. Comput. Vis. pp. 568–578 (2021)
51. Yao, T., Li, Y., Pan, Y., Wang, Y., Zhang, X.P., Mei, T.: Dual vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
52. Zeng, W., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11101–11111 (2022)
53. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. Int. Conf. Learn. Represent. (2023)
54. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: IEEE Int. Conf. Image Process. pp. 276–280. IEEE (2020)
55. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 72–80 (2021)
56. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. IEEE Trans. Intell. Transp. Syst. (2023)
57. Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R.: Delivering arbitrary-modal semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1136–1147 (2023)
58. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3221 (2017)
59. Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., Yue, X.: Meta-transformer: A unified framework for multimodal learning. arXiv preprint arXiv:2307.10802 (2023)
60. Zheng, A., Zhang, Y., Zhang, X., Qi, X., Sun, J.: Progressive end-to-end object detection in crowded scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 857–866 (2022)
61. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. Int. Conf. Learn. Represent. (2024)
62. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Int. Conf. Learn. Represent. (2020)
63. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: Int. Conf. Comput. Vis. pp. 6748–6758 (2023)