

# Detection of Micromobility Vehicles in Urban Traffic Videos

Khalil Sabri, Célia Djilali, Guillaume-Alexandre Bilodeau, Nicolas Saunier  
Polytechnique Montréal  
Montréal, Canada  
{khalil.sabri,celia.djilali,gabilodeau,nicolas.saunier}@polymtl.ca

Wassim Bouachir  
University of Québec (TÉLUQ)  
Montréal, Canada  
wassim.bouachir@teluq.ca

**Abstract**—Urban traffic environments present unique challenges for object detection, particularly with the increasing presence of micromobility vehicles like e-scooters and bikes. To address this object detection problem, this work introduces an adapted detection model that combines the accuracy and speed of single-frame object detection with the richer features offered by video object detection frameworks. This is done by applying aggregated feature maps from consecutive frames processed through motion flow to the YOLOX architecture. This fusion brings a temporal perspective to YOLOX detection abilities, allowing for a better understanding of urban mobility patterns and substantially improving detection reliability. Tested on a custom dataset curated for urban micromobility scenarios, our model showcases substantial improvement over existing state-of-the-art methods, demonstrating the need to consider spatio-temporal information for detecting such small and thin objects. Our approach enhances detection in challenging conditions, including occlusions, ensuring temporal consistency, and effectively mitigating motion blur.

**Index Terms**—Urban Traffic, Micro-Mobility Detection, Object Detection, YOLO, Video Object Detection, Autonomous Vehicles, Urban Transportation Safety

## I. INTRODUCTION

In recent years, urban transportation has undergone a significant transformation with the emergence of micromobility solutions, in particular electric version of bikes, scooters, and skateboards. These modes of transport, renowned for their agility and environmental benefits, are reshaping short-distance commutes in bustling cities. However, the integration of these vehicles into the urban landscape is not without challenges. Their unique manoeuvrability and smaller footprints, while advantageous for users, complicate detection amidst heavy traffic, elevating safety risks.

The objective of this work is to design a robust micromobility vehicles (MMV) detection system, a critical component in ensuring the safety of both riders and surrounding traffic, particularly in the era of automated vehicles. As well, counting MMVs, their origins and destinations is an important component of efficient city planning. As micromobility solutions become increasingly prevalent in urban landscapes, the integration of such detection systems is paramount, not only for current traffic dynamics but also as a foundational technology for the seamless operation of future automated transportation systems.

This paper presents a novel detection model, FGFA-YOLOX, that synthesizes the strengths of image-based and video-based object detection methodologies. Our approach leverages the rapid and efficient image analysis capability of the YOLOX framework [6], integrating it with the temporal context consideration of video object detection systems ([17]–[19]). This fusion aims to enhance detection consistency and accuracy in urban traffic scenarios, addressing the unique challenges posed by micromobility vehicles. This is done by applying aggregated feature maps from consecutive frames processed through motion flow to the YOLOX architecture. With this strategy, our method can benefit from the large number of readily available pre-trained models for YOLOX. Aggregation of feature maps then enhances the capabilities of the detector, thanks to the combination of several of object views, with some more informative than others.

To rigorously evaluate the performance of our proposed model against state-of-the-art (SOTA) methods, we constructed a new custom dataset focused on micromobility scenarios. Our findings indicate that our proposed model achieves superior performance compared to SOTA methods, demonstrating the need to consider spatio-temporal information for detecting such small and thin objects. Our contributions can be summarized as follows.

- 1) We propose a novel video object detection architecture, FGFA-YOLOX, adapted for MMV and taking advantage from both the capabilities of single-frame object detection (accuracy and speed, several available pre-trained models) and the richer feature representation of video object detectors;
- 2) We constructed a new dataset that we make publicly available for the evaluation of MMV object detectors, along with our source code that we provide to ensure reproducibility and reusability. Our model weights, code, and dataset are publicly available for further research and replication of our results at [https://github.com/sabrikhalil/Micro\\_Mobility\\_Detection](https://github.com/sabrikhalil/Micro_Mobility_Detection).

## II. RELATED WORK

Few works have considered micromobility vehicle (MMV) detection. The study by Apurv et al. [1] introduced an approach for identifying e-scooter riders using a system comprised of two distinct modules. The first module employs a

pre-trained YoloV3 model [14] for effective initial detection of pedestrians. Following this, the second module comes into play, which involves expanding the bounding boxes around the detected individuals to encompass the e-scooter along with the rider. This expansion is essential to ensure the entire vehicle, including both the rider and the e-scooter, is captured. The process is further refined using a MobileNetV2 classifier [15], trained on a specialized dataset, to distinguish whether a person is with or without an e-scooter. This two-module system permits to accurately identify e-scooter riders. Building on this previous work, Gilroy et al. [7] extended the research for overcoming occlusion challenges in urban environments. They modify the YoloV3 architecture, tailoring it to detect partially visible e-scooter riders more reliably. This enhancement significantly improves the detection rates in scenarios where riders are obscured.

In contrast to these two previous works that focus on detecting a single vehicle category, e-scooter, our research aims to include a variety of MMVs. We thus aim to establish a more versatile framework. Our goal is to enhance MMV detection capabilities to not only address occlusion challenges, but also to incorporate resilience against motion blur, a frequent issue in rapidly moving vehicles, and to ensure consistent detection across video frames.

In single-frame object detection, two primary categories emerge: two-stage and one-stage models. The two-stage models, exemplified by R-FCN [4], adopt a sequential approach. Initially, region proposals are generated and subsequently they are classified into different object categories. This intricate process, while delivering precision, tends to be more computationally intensive, leading to slower inference times compared to one-stage models. One-stage models, like YOLO [13] and its iterations like YOLOX [6], streamline the detection process. YOLO considers object detection as a single regression problem, directly moving from image pixels to bounding box coordinates and class probabilities. Nevertheless, for all single frame approaches, their frame-by-frame nature might lead to detection inconsistency in video due to challenges like motion blur and occlusions, highlighting the need for models that maintain temporal consistency.

In the evolving landscape of video object detection, Deep Feature Flow (DFF) [19] marked a significant progress. DFF emerged as a response to the computational demand of traditional methods. By innovatively employing keyframe feature extraction and utilizing optical flow, DFF transfers these features to adjacent non-key frames. This methodology not only accelerates the detection process but also alleviates the computational burden, making it a pivotal development for more efficient video object detection. Building on DFF emphasis on efficiency, Flow-guided Feature Aggregation (FGFA) [18] took a step further, enhancing the quality of the features. FGFA aggregates and adaptively weights deformed feature maps from neighbouring frames. This process, aided by optical flow, improves accuracy by integrating relevant data across frames, addressing challenges like motion blur and occlusion. Similarly, Sequence Level Semantics Aggregation

(SELSA) [17] was proposed with a focus on the semantic relationships between objects across frames. SELSA analyzes and aggregates features based on semantic similarities, leading to more contextually aware and precise object detection.

### III. METHODOLOGY

#### A. Motivation

While one-stage single-frame detection models like YOLOX [13] are efficient, they struggle with issues specific to videos, such as inconsistent detections over frames, motion blur, and occlusions. On the other hand, video object detection models that consider multiple frames provide better temporal consistency, but often lack the speed, the modern feature extraction strategies and the availability of pre-trained weights of one-stage models. To bridge this gap, we propose to combine the speed and accuracy of YOLOX [6] with the temporal coherence of Flow-guided Feature Aggregation (FGFA) [18], aiming for a balanced solution in video object detection. We name our method FGFA-YOLOX. Figure 1 provides an overview of our proposed detection framework, showcasing the enhancement of the current frame feature map with motion-adjusted neighbour frames and subsequent processing through the YOLOX architecture neck and head for effective detection.

#### B. Problem Statement

Detecting micromobility vehicles (MMV) like e-scooters, bicycles, and skateboards in urban traffic involves challenges due to their size, movements, and the potential for occlusions. Given a sequence of video frames  $F = \{f_{t-N}, \dots, f_t, \dots, f_{t+N}\}$ , where  $f_t$  is the current frame at time  $t$  and  $N$  indicates the number of contextual frames in the past and in the future, our goal is to accurately detect MMV in  $f_t$ . The goal is to find a function  $D$  that, applied to the current frame  $f_t$  and an aggregated feature map  $\mathcal{G}$  derived from both  $f_t$  and its contextual frames, yields detections defined by bounding boxes  $B$ , class labels  $L$ , and confidence scores  $S$ :

$$D(f_t, \mathcal{G}) \rightarrow \{(B, L, S)\}. \quad (1)$$

#### C. Integrated Detection Model Framework

**Feature Extraction:** As shown in Figure 1, the first step of our approach is to perform feature extraction. The CSPDarknet backbone  $\mathcal{B}$  extracts spatial features  $\mathcal{S}_t$  from the current frame  $f_t$  and  $\mathcal{S}_{\text{context}}$  from the neighbouring contextual frames, defined as:

$$\mathcal{S}_t = \mathcal{B}(f_t), \quad (2)$$

and

$$\mathcal{S}_{\text{context}} = \bigcup_{i=t-N, i \neq t}^{t+N} \mathcal{B}(f_i). \quad (3)$$

The CSPDarknet backbone, selected for feature extraction in our model, was shown to be very efficient and capable by Wang et al. [16]. This architecture is efficient in optimizing gradient flow and reducing computational load, making it ideal

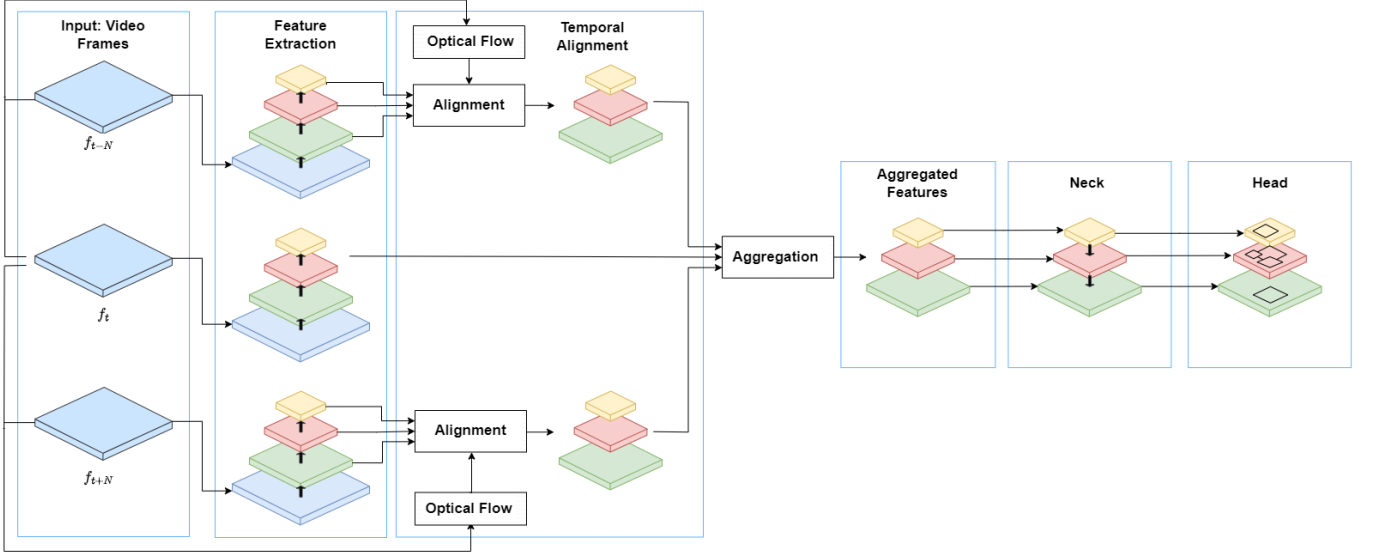


Fig. 1: FGFA-YOLOX detection framework overview. Features from input frames are first extracted with the backbone of YOLOX. The optical flows of the current frame with past and future frames (neighbour frames) are also computed. For temporal aggregation, the motion-adjusted features of the neighbour frames are aggregated with those of the current frame. The aggregated features are then processes through the YOLOX architecture neck and head for detection.

for detecting objects of varying sizes, particularly micromobility vehicles (MMV) in urban environments. For an input image, this backbone produces feature maps with dimensions  $H \times W \times C$ , where  $H$  and  $W$  represent the height and width of the reduced spatial resolution of the output feature maps, and  $C$  symbolizes the channel depth, which is  $C_1$ ,  $C_2$ , or  $C_3$  for each scale, adapting to the pyramid feature network requirements. This downsampling is essential for processing high-resolution inputs efficiently while retaining significant spatial features necessary for accurate object detection.

**Temporal Alignment and Aggregation:** As shown in Figure 1, the second step is feature aggregation. A motion estimation function  $\mathcal{M}$  aligns context frame features  $\mathcal{S}_{\text{context}}$  with the current frame feature  $\mathcal{S}_t$ . This alignment ensures that the features are spatially coherent and results in features:

$$\mathcal{S}_{\text{aligned}} = \mathcal{M}(\mathcal{S}_{\text{context}}, \mathcal{S}_t). \quad (4)$$

FlowNetSimple [5] is used to implement the motion estimation function  $\mathcal{M}$  to align feature maps from context frames to the current frame feature map. This process aligns  $\mathcal{S}_{\text{context}}$ , with dimensions  $H \times W \times C$ , to the spatial configuration of  $\mathcal{S}_t$  to closely mirror the current frame feature map in both spatial and channel dimensions, effectively capturing the estimated features at time  $t$ . This alignment enables a more accurate representation of the scene at the current moment, laying the groundwork for subsequent feature aggregation to further enhance detection capabilities in evolving urban environments.

These aligned features are aggregated with  $\mathcal{S}_t$  via the aggregation function  $\mathcal{A}$ , forming an enriched feature map  $\mathcal{G}$  that captures both spatial and temporal characteristics:

$$\mathcal{G} = \mathcal{A}(\mathcal{S}_{\text{aligned}}, \mathcal{S}_t). \quad (5)$$

In our method, the aggregation function  $\mathcal{A}$  employs concatenation and convolutions for merging  $\mathcal{S}_{\text{aligned}}$  and  $\mathcal{S}_t$ . This operation is realized through a concatenation step, formulated as:

$$\mathcal{S}_{\text{stacked}} = \text{Concat}(\mathcal{S}_{\text{aligned}}, \mathcal{S}_t), \quad (6)$$

that effectively doubles the channel size of the input features, preparing them for the subsequent convolution process. The convolution, employing a  $3 \times 3$  kernel, is designed to integrate and refine the concatenated features, resulting in:

$$\mathcal{G} = \text{Conv}_{3 \times 3}(\mathcal{S}_{\text{stacked}}), \quad (7)$$

where  $\mathcal{G}$  is the output feature map with dimensions  $H \times W \times C$ , maintaining the original spatial dimensions while encapsulating an enriched representation of both spatial and temporal information.

**Detection and Classification:** Finally, as shown in Figure 1, detection and classification is performed through the neck and head. After aggregating features to obtain  $\mathcal{G}$ , the YOLOX-PAFPN neck  $\mathcal{N}$  is employed. The 'PAFPN' in YOLOXPAFPN denotes Path Aggregation Feature Pyramid Network, which enhances multi-scale feature fusion by leveraging the Path Aggregation Network (PAN) for efficient optimization [12]. This configuration refines the feature processing across scales, optimally preparing them for the detection task. The YOLOX head  $\mathcal{H}$  processes the enriched feature map to detect Micromobility Vehicles (MMV) within the current frame  $f_t$ . This sequential application of the neck and head on  $\mathcal{G}$  is given by:

$$(B, L, S) = D(f_t, \mathcal{G}) = \mathcal{H} \circ \mathcal{N}(\mathcal{G}). \quad (8)$$

In more details, our method employs a feature pyramid network (FPN) [10] in the form of YOLOXPAFPN for integrating the multi-scale feature maps. This approach ensures good detection across various object sizes, crucial for smaller targets requiring high-resolution recognition. The process involves merging aggregated feature maps  $\mathcal{G}_{C1}$ ,  $\mathcal{G}_{C2}$ , and  $\mathcal{G}_{C3}$ , each indicative of a unique scale within the detection framework:

$$\mathcal{N}(\mathcal{G}) = \text{FPN}(\mathcal{G}_{C1}, \mathcal{G}_{C2}, \mathcal{G}_{C3}) \quad (9)$$

This integration yields  $\mathcal{N}(\mathcal{G})$ , a composite feature map of  $H \times W \times C$ , standardizing the output to facilitate precise, scale-invariant object detection. Through this pyramid approach, the model effectively consolidates spatial and scale data, enhancing detection accuracy in varied urban environments.

The YOLOX head,  $\mathcal{H}$ , which is the concluding component in our detection framework, utilizes the feature map  $\mathcal{N}(\mathcal{G})$  to delineate the location of objects and classify them. It conducts a detailed analysis on  $\mathcal{N}(\mathcal{G})$ , outputting a set of detections  $\mathcal{D}$ , each defined by a bounding box  $B$ , a class label  $L$ , and a confidence score  $S$  as presented in equation 8.

This operation is key to identifying and classifying MMV accurately, incorporating confidence scores  $S$  to gauge the reliability of each detection. By handling  $\mathcal{N}(\mathcal{G})$  from the YOLOXPAFPN neck, the head ensures detailed detection across varying object sizes, essential for monitoring dynamic urban scenes. This approach maintains consistency with our problem statement, underlining the necessity of combining current frame analysis with contextual features for robust and accurate MMV detection.

#### IV. EXPERIMENTS

This section outlines the experimental setup and the evaluation of our proposed model, including the dataset used, evaluation metrics, implementation details, and comparison with state-of-the-art (SOTA) models.

##### A. Dataset

1) *Dataset Collection and Construction:* A custom dataset, named PolyMMV, was developed to address the detection of MMV, given the scarcity of suitable existing datasets. Sourced from a variety of online public video hosting platforms, this dataset aims to mirror the diversity of real-world urban micro-mobility, covering bicycles, skateboards, and electric scooters as primary classes.

For research reproducibility and further exploration, the dataset, including annotations in YOLO, COCO, and VOC formats, is available at our GitHub repository: [https://github.com/sabrikhalil/Micro\\_Mobility\\_Detection](https://github.com/sabrikhalil/Micro_Mobility_Detection). This initiative supports the advancement of urban MMV detection research. Figure 2 presents samples of annotated images from the training set. These examples, randomly chosen and representing various classes, showcase the real-world diversity in object sizes, positions, lighting conditions, and occasional occlusions. Some

images also feature objects from multiple classes, adding to the dataset complexity.

2) *Annotation Process and Dataset Characteristics:* The videos were annotated using the Computer Vision Annotation Tool (CVAT). The annotation process involved labelling bicycles, skateboards, and electric scooters with bounding boxes. Figure 3 shows the characteristics of our dataset. The training set contains 80 videos, and the test set comprises 25 videos. As seen in Figure 3, the training data includes around 6000 bounding boxes for bicycles, and approximately 5000 each for electric scooters and skateboards. This balanced distribution facilitates the training of the model across different MMV classes. The diverse sizes and positions of the bounding boxes, representative of real-world scenarios, are also visible in the figure.

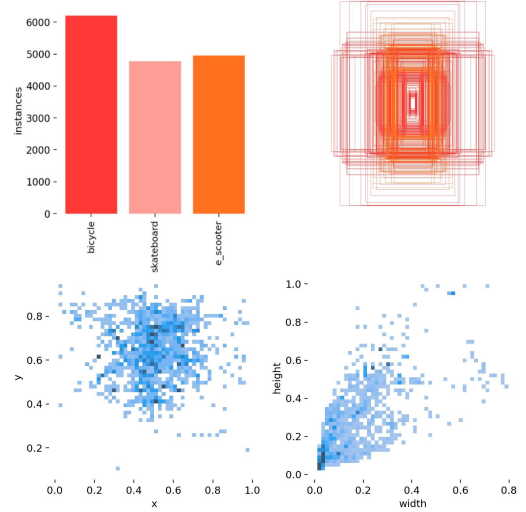


Fig. 3: Characteristics of the training dataset of PolyMMV. Top left: the number of instances distribution of bicycles, skateboards, and e-scooters, top right: illustration of the size distribution of bounding boxes, bottom left and right depict the scatter plots of normalized bounding box positions and sizes, respectively.

In contrast, the test set is designed to evaluate the model generalization capabilities and includes 4000 instances of bicycles, 2500 skateboards, and 2000 electric scooters.

##### B. Evaluation Metrics

For the evaluation of our model, we used the mean Average Precision (mAP) and mAP@50 metrics. The mAP provides an overall effectiveness of the model by averaging the precision across different recall levels and object categories, and mAP@50 specifically considers detections as correct if they have an Intersection over Union (IoU) of more than 50% with the ground-truth. AP is similar to the mAP, but used for each individual object class. These metrics are particularly relevant in object detection tasks to measure the accuracy of the model in identifying and localizing objects correctly. Our evaluation metrics align with the standards set by the COCO

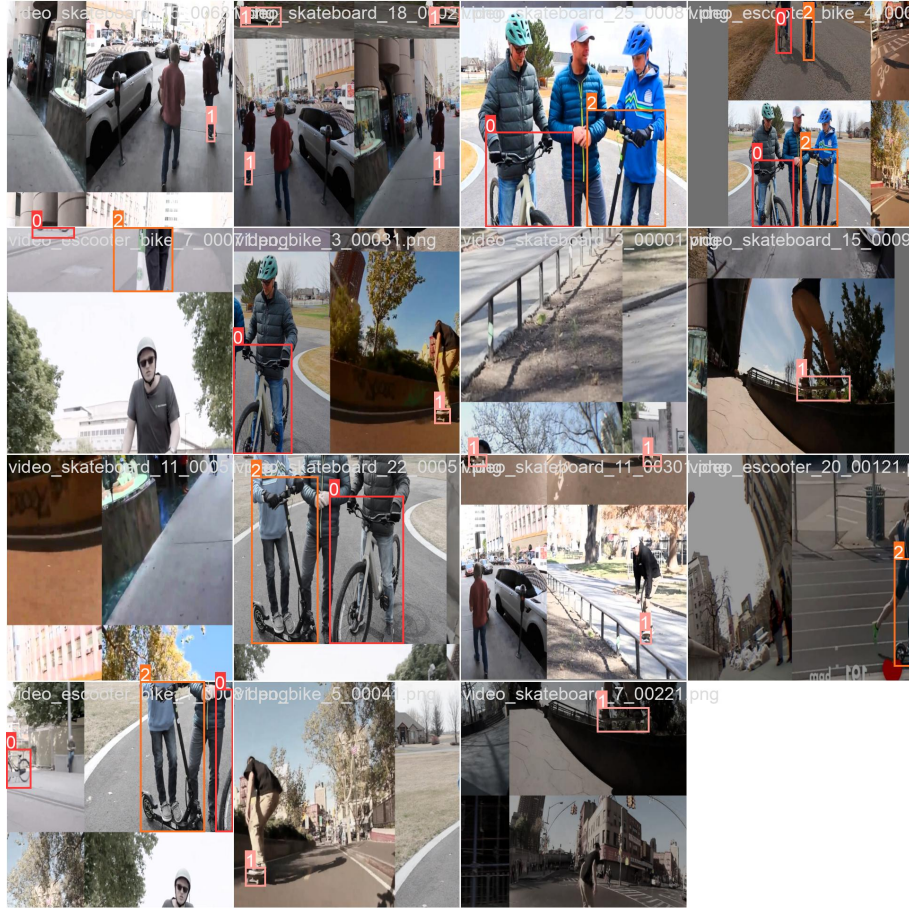


Fig. 2: Examples of annotated images in PolyMMV. Class 0: Bicycles, Class 1: Skateboards, Class 2: Electric Scooters

evaluation protocols, which are widely recognized for their comprehensiveness in assessing object detection performance [11].

### C. Implementation Details and Model Training

Our models were implemented using Python and the PyTorch framework, alongside mmtracking [3] for video object detection and OpenCV [2] for video data handling. We integrated YOLOX into the FGFA framework, utilizing components that were pre-trained on specific datasets to enhance their robustness and effectiveness. The YOLOX backbone used for object detection, was pre-trained on the COCO dataset [11]. Its configuration, with a deepen factor of 1.33 and a widen factor of 1.25, effectively enhances the network ability to capture intricate details without excessive computational demands. This is crucial for real-time video analysis. The input image size is  $640 \times 640$ , giving values  $C_1 = 320$ ,  $C_2 = 640$  and  $C_3 = 1280$ . The FlowNetSimple [5], [8] module is used for motion estimation and was pre-trained on the Flying Chairs dataset [5], as described in the original FlowNet study.

Our FGFA-YOLOX model and the DFF-YOLOX variant underwent fine-tuning from COCO pre-trained weights using an SGD optimizer with an initial learning rate of 0.0001, momentum of 0.9, and weight decay of 0.00001 over

three epochs, incorporating a warm-up phase during the first 500 iterations. Similarly, RFCN-based methods (DFF-RFCN, FGFA-RFCN, SELSA-RFCN) were initially trained with an R-FCN detector on COCO video data, with subsequent fine-tuning for our MMV detection task that adjusted the learning rate to 0.01, alongside comparable momentum and weight decay settings. For YOLOv8, originally designed for single-image object detection and also pre-trained on the COCO dataset, we adapted it for video detection by sampling at 10 frames per second to avoid overfitting from frame redundancy and utilized dropout techniques to improve generalization. This multifaceted approach to training, including specific learning rate adjustments and dataset-oriented refinements for each model, was crucial in enhancing detection capabilities across diverse urban environments.

### D. Comparison with State-of-the-Art Object detectors

Table I presents a performance comparison between our proposed detector FGFA-YOLOX and various SOTA object detection models, highlighting its superior performance, achieving the highest scores in mAP and mAP@50. It particularly improves in detecting skateboards, a challenging category due to their smaller size, frequent occlusion by the rider's feet, and motion blur as evidenced in Figure 4B. Bicycles also

Models	mAP	mAP@50	AP-bicycle	AP-skateboard	AP-scooter	Inference Time per Frame
DFF - RFCN [19]	27.9	57.6	33.8	8.2	41.7	41.4
FGFA - RFCN [18]	31.2	61.1	38.9	10.0	44.7	418.1
SELSA - RFCN [17]	31.0	62.4	36.6	11.6	44.7	317.0
YOLOv8 [9]	34.5	64.2	39.2	17.7	46.7	<b>34.2</b>
<b>FGFA - YOLOX (ours)</b>	<b>38.6</b>	<b>69.4</b>	<b>45.0</b>	<b>23.2</b>	<b>47.6</b>	329.7

TABLE I: Comparison of model performances and inference times on PolyMMV dataset. mAP and AP are in percentage. Inference time are in milliseconds. Best scores in **bold** font.

saw a considerable increase in detection accuracy, benefiting from the model’s robust feature extraction and aggregation capabilities. In contrast, e-scooters, which tend to move slower and exhibit more consistent movement patterns, showed less improvement. This variance underscores the model’s adaptability to different object characteristics within urban traffic scenarios.

TABLE II: Confusion matrix for the detection of micromobility vehicles on PolyMMV, assuming an IoU threshold of 0.5.

Actual Class	Predicted Class			
	Bicycle	Skateboard	E-scooter	Background
<b>Bicycle</b>	0.83	0	0	0.17
<b>Skateboard</b>	0	0.44	0	0.56
<b>E-scooter</b>	0.02	0	0.66	0.32
<b>Background</b>	0.57	0.26	0.17	0

The confusion matrix for our FGFA-YOLOX confirms our previous observations. The model excels in separating classes by learning from several frames. Specifically, 44% of skateboards are correctly detected while this number is 31% for FGFA-RFCN and 35% for YOLOv8 [9], representing a significant improvement. However, there are areas for improvement, particularly in fine-tuning IoU thresholds to better balance the detection of micromobility vehicles against complex urban backgrounds. This adjustment is aimed at reducing the higher incidence of false positives observed, enhancing the model’s precision. Further diversifying our training dataset will also support this goal, enabling the model to more accurately recognize electric scooters, bicycles, and skateboards in a variety of urban scenarios.

Table I also illustrates the inference times of various detection models. Our FGFA-YOLOX model notably achieves faster inference times compared to FGFA-RFCN, but shows that using several frame for detecting is more costly than for example YOLOv8.

#### E. Model Performance in Various Scenarios

To illustrate the performance of FGFA-YOLOX, we will discuss in the following sample qualitative results in three complex urban traffic scenarios: occlusion, motion blur, and temporal consistency. The figure 4 below contrasts our model performance with that of single-frame detection models like YOLOv8 and video-based models like FGFA-RFCN, showcasing our approach robustness in handling these challenges.

**Occlusion Handling:** In dense urban environments, vehicles often become partially occluded. Our model excels in such scenarios, effectively detecting MMVs despite occlusions.

As depicted in figure 4A, our model successfully identifies vehicles that are partially hidden, a task where the YOLOv8 model struggles due to its reliance on single-frame analysis.

**Motion Blur:** Figure 4B illustrates our FGFA-YOLOX ability to handle motion blur caused by rapid movements, which is a significant challenge for single-frame models, like YOLOv8, that do not utilize past frames for continuous object identification and can be affected by motion blur. Unlike YOLOv8, our model mitigates the effects of motion blur, ensuring more detection of fast-moving objects, particularly skateboards.

**Improved Temporal Consistency:** Figure 4C demonstrates that our model enhances temporal consistency. While traditional video object detection models may show variability in detections across frames, our model maintains stable and accurate detection, important for real-time monitoring and autonomous navigation in urban traffic.

The combined strengths of our model, merging the precision of single-frame detection with the comprehensive context of video object detection, offer a significant advancement in addressing the diverse challenges of urban traffic conditions for MMV detection.

## V. CONCLUSION

This paper introduces a novel detection model that leverages both single-frame precision and video object detection capabilities for accurately identifying MMV in urban environments. Our approach, validated on a new MMV dataset, demonstrates significant improvements over existing methods by incorporating spatio-temporal information for enhanced detection performance.

Our research opens avenues for exploring attention mechanisms and innovative architectures for feature aggregation. Integrating attention mechanisms into the model could refine its focus on relevant features, improving occlusion handling and detection of small or partially visible vehicles.

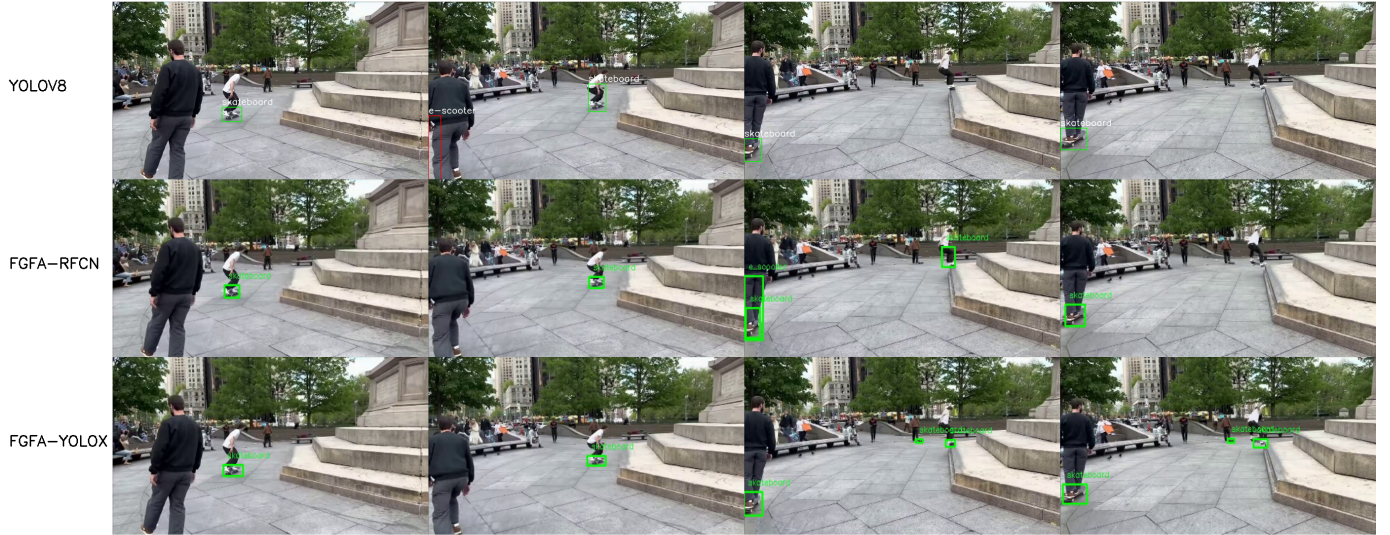
## REFERENCES

- [1] Kumar Apurv, Renran Tian, and Rini Sherony. Detection of e-scooter riders in naturalistic scenes. *arXiv preprint arXiv:2111.14060*, 2021.
- [2] G. Bradski. The OpenCV Library. *Dr. Dobbs’ Journal of Software Tools*, 2000.
- [3] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. <https://github.com/open-mmlab/mtracking>, 2020.
- [4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

A) Occlusion



B) Motion Blur



C) Temporal Consistency

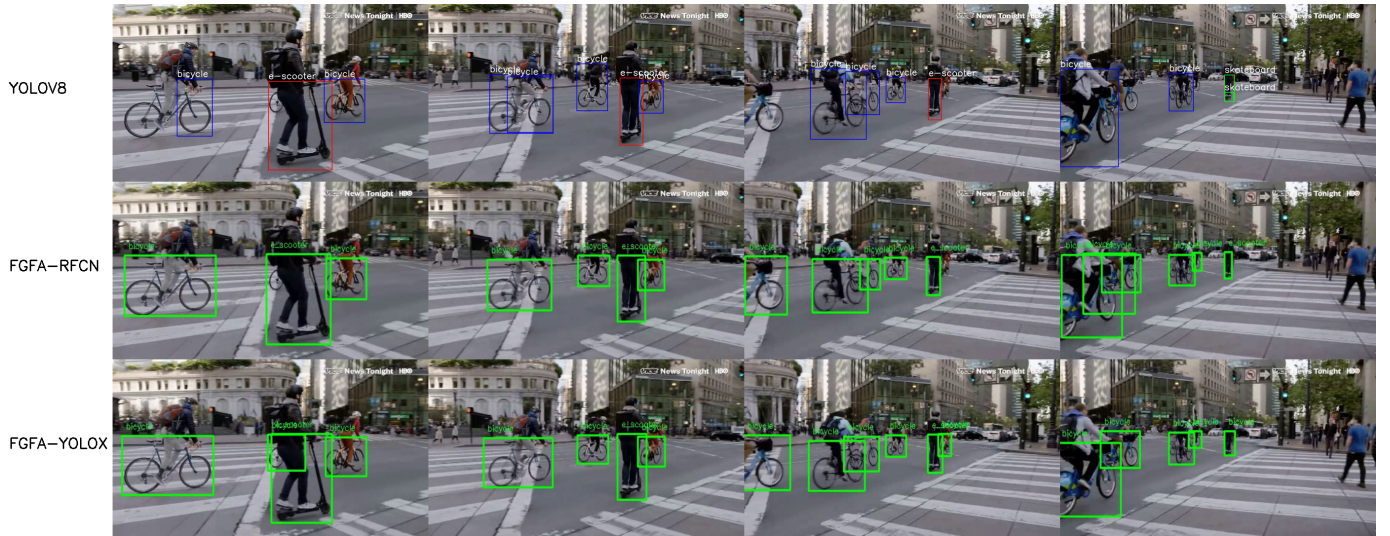


Fig. 4: Comparative analysis of model performance in various scenarios. A) Occlusion, B) Motion blur, and C) Temporal consistency.

- [5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.
- [6] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [7] Shane Gilroy, Darragh Mullins, Edward Jones, Ashkan Parsi, and Martin Glavin. E-scooter rider detection and classification in dense urban environments. *Results in Engineering*, 16:100677, 2022.
- [8] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.
- [9] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
- [10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [12] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.
- [13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [14] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [16] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.
- [17] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9217–9225, 2019.
- [18] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 408–417, 2017.
- [19] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2349–2358, 2017.