Road Object Detection in Fish-Eye Cameras

RE6124019
Ming-Hsuan Wu
Institute of Data Science
RE6124019@gs.ncku.edu.tw

RE6124027
Wen-Hai Tseng
Institute of Data Science
RE6124027@gs.ncku.edu.tw

RE6121011
Jen-Lung Hsu
Institute of Data Science
RE6121011@gs.ncku.edu.tw

N96124153
Hsu-Chun-Che
Institute of Engineering Science
N96124153@gs.ncku.edu.tw

Abstract

This study aims to address the object recognition challenge in circular fisheye images, a task that presents significant difficulties in the field of computer vision. We participated in the AI City Workshop competition held at CVPR, specifically focusing on road target detection using fisheye cameras. We propose a multiscale and multiresolution method that combines geometric calibration, transformations, and deep learning models (such as the YOLO series) to tackle this problem. Our research evaluates the performance on the FishEye8K and FishEye1Keval datasets, demonstrating competitive object recognition capabilities in circular fisheye imagery.

1. Introduction

The surveillance of vehicular movement is a crucial aspect of modern urban transportation systems, providing essential data on traffic dynamics, incident identification, and facilitating effective traffic management. Traditional cameras, with their restricted fields of view (FoV), often fail to offer a comprehensive perspective of roadways and intersections, thereby necessitating the deployment of multiple cameras for adequate coverage. Fisheye lenses, capable of capturing wide-ranging, omnidirectional vistas with a single camera, have recently gained traction, offering an attractive alternative for traffic surveillance tasks [25].

Nonetheless, fisheye cameras pose their unique set of challenges: they generate distorted, curved imagery that requires specialized image processing techniques for rectification and dewarping [6, 13]. This intricate task has hindered the widespread incorporation of fisheye cameras in traffic surveillance systems. Moreover, there is a dearth of publicly available datasets specifically tailored for fisheye road object detection, which hampers the development and evaluation of state-of-the-art computer vision algorithms in this domain.

To surmount these hurdles, we propose a comprehensive research project aimed at developing an efficient traffic surveillance system employing fisheye cameras and cutting-edge deep learning object detection methodologies. Our project will leverage the recently launched FishEye8K [8] and FishEye1Keval datasets, encompassing a diverse range of traffic scenarios, illumination conditions, and viewing angles of five road object categories at varying scales. These datasets will serve as a priceless resource for training and evaluating our proposed object detection models.

In addition to utilizing the FishEye8K and FishEye1Keval datasets, we harness panoramic road images to train a robust teacher model. By employing knowledge distillation techniques, we perform semi-automatic labeling of unlabeled fisheye images, thereby creating a student model specifically tailored for object recognition in fisheye imagery. Central to our approach is the integration of the Parallel Residual Bi-Fusion (PRB) technique into the YOLO model, which enhances fisheye image quality and mitigates the impact of fisheye distortions on object recognition, ultimately boosting the model's performance.

2. Related Work

2.1. Feature-Based Approaches

These methods typically utilize local feature descriptors such as SIFT [18], SURF [1], and ORB [21] for object detection and matching in images. However, in the case of fisheye imagery, traditional feature detection and descriptor matching may be affected due to distortion and deformation [5].

2.2. Deep Learning Methods

In recent years, deep learning has achieved significant success in object recognition tasks [19]. For circular fisheye images, convolutional neural networks (CNNs) can be employed for feature extraction and object recognition. Leveraging the powerful capabilities of CNNs, discriminative features can be learned directly from raw fisheye images.

2.3. Geometric Calibration and Transformations

The unique projection characteristics of fisheye images allow for geometric calibration and transformations to convert circular fisheye images into regular perspective projection images [6, 13]. Following this transformation, traditional object recognition methods or deep learning approaches can be applied for object detection.

2.4. Multiscale and Multiresolution Methods

Features extracted from images at different scales and resolutions contribute to effective learning in deep neural network models [3,15]. Therefore, adopting multi-scale and multi-resolution techniques can enhance target recognition accuracy and improve efficiency in fisheye image processing.

2.5. YOLO Series Object Detection Models

The YOLO (You Only Look Once) series has emerged as a prominent family of single-stage object detection models, renowned for their exceptional inference speed and accuracy. Unlike two-stage detectors, YOLO models directly predict object categories and bounding boxes in a single pass, making them highly suitable for real-time applications such as autonomous driving [14, 20, 22, 23]. Over the years, the YOLO series has undergone continuous improvements, with each new version introducing architectural enhancements and training strategies to further boost performance. Notable versions of the YOLO series include YOLOv7 [22], YOLOv8 [20], YOLOv6 3.0 [14], and the most recent addition, YOLOv9 [23]. These models showcase remarkable inference speed while maintaining high accuracy, making them ideal for resource-constrained environments. The architectures of these models are carefully designed to handle multi-scale processing, enabling them to effectively detect objects of various sizes [14, 20, 22, 23]. By leveraging techniques such as feature pyramid networks, anchor-free detection, and advanced backbone networks, the YOLO series continues to push the boundaries of real-time object detection performance.

3. System framework

We anticipate propose a multiscale and multiresolution approach that combines geometric calibration and transformations with deep learning models (such as the YOLO series) to address the object recognition challenge in circular fisheye images.

Firstly, we employ models from the YOLO series for multiscale object recognition. By adjusting the model architecture and enhancing the images, we improve the model's detection capabilities across different scales and resolutions. Secondly, we conduct specialized model training for different objects and scenarios. During the model embedding process, we integrate these specialized models to enhance the model's accuracy in object recognition.

Upon observing the results, we found that the evaluation of photos in nighttime scenes performed poorly. This may be due to the lower resolution of photos in nighttime scenes and the relatively small number of samples. Therefore, our next strategy is to separately train models in daytime and nighttime scenes, as illustrated in Figure 1, followed by experimental evaluations. However, we have not yet begun this part of the experiment.

Figure 1. Fisheye Images

3.1. The Proposed Adapting Object Detection to Fisheye Cameras

In Figure 2, we illustrate the training process. We employ a Teacher model to train a powerful model using panoramic images and image augmentation. The Teacher model generates pseudo-labels, and we train the Student model using these pseudo-labels. The improvement in model mAP after training indicates the usefulness of pseudo-labels. The confidence filter determines which pseudo-labels we should retain, leading to the next round of training.

3.2. PRB-FPN with YOLO

The foundation of our object detection system is built upon the synergy between the YOLO object detection model and the Parallel Residual Bi-Fusion (PRB) [4] technique. In this section, we elucidate the mechanics of these core components, discussing their pivotal roles and inherent benefits within our system.

YOLO: YOLO stands as a vanguard in real-time object detection, showcasing remarkable inference speed—a trait

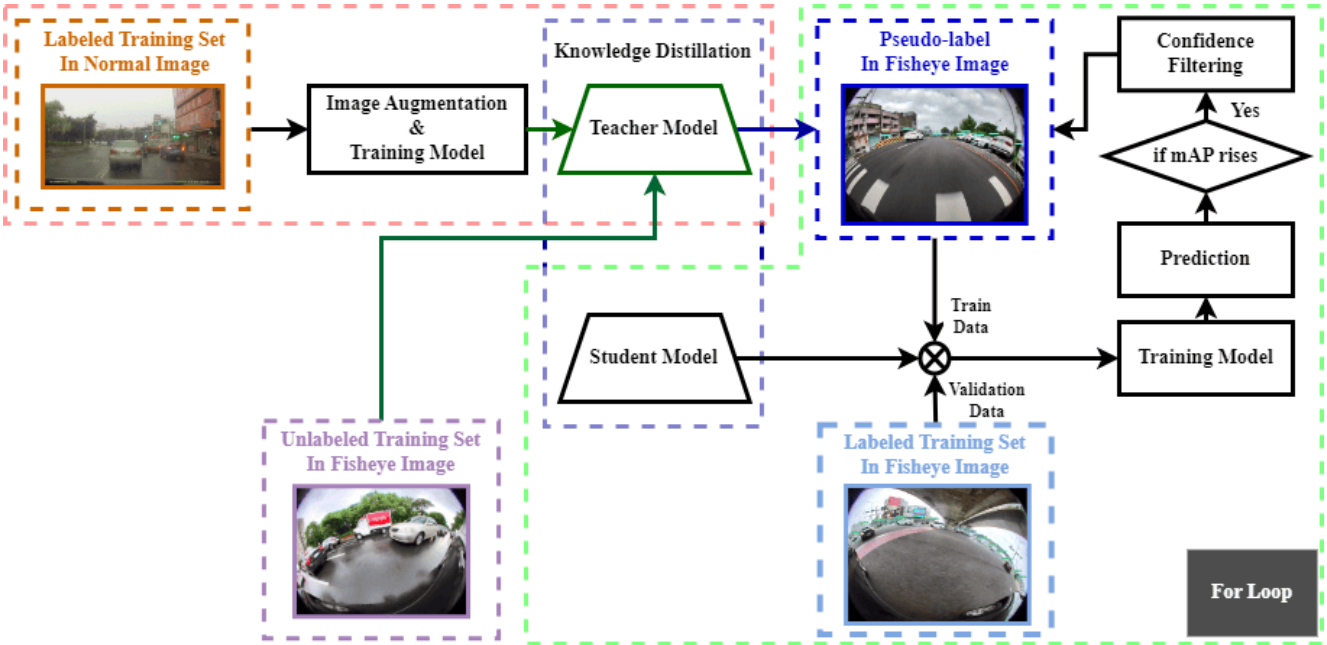

Figure 2. Training Model System Flowchart

indispensable for real-world applications like autonomous driving. In our experiments, we primarily employ YOLOv6 3.0, YOLOv6 3.0+PRB, YOLOv7+PRB, YOLOv9, and YOLOv9+PRB to conduct a comparative analysis of their performance on the FishEye8K dataset.

PRB-FPN: The PRB-FPN [4], a fundamental component of the YOLO-PRB model, is crafted to build a feature pyramid network [7], [16], targeting object detection across varied scales.

Expanding upon the traditional Feature Pyramid Network (FPN) framework, PRB-FPN [3] deploys a bidirectional fusion strategy. This strategy amalgamates high-level features with their low-level counterparts and vice versa, ensuring a comprehensive capture of multiscale information. This in turn augments the precision of object detection. A key feature embedded within is the integration of residual connections [10], termed Re-Core. By mitigating potential information loss and countering the gradient vanishing problem, these connections bolster the model's performance in object detection tasks. A graphical depiction of the Bi-fusion Module is provided in Figure 3.

Due to the abundance of training and testing data and considering the computational resources and time costs involved, we optimized the PRB structure by truncating the core blocks from three to a more streamlined two. Empirical evidence indicates that this strategic reduction slashes the model size by an impressive 52%, with a minimal decrement in mAP, measured at 3%. This approach enables rapid completion of the student model without significant performance degradation. Post this architecture finalization for both Teacher and Student Models, our training approach pivots on leveraging knowledge distillation [11], [9], [17] in tandem with the semi-pseudo-label [12], [2] semi-supervised learning techniques. The intricate details of this combined approach are illustrated in Figure 2.

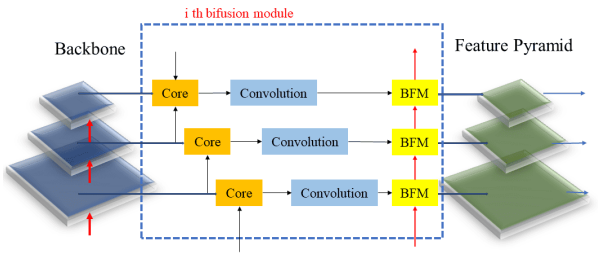

Figure 3. Parallel Residual Bi-Fusion Model Structure

3.3. Pseudo-Label-Guided Knowledge Distillation

In the initial step, we train the YOLO+PRB model using the training dataset (containing panoramic images and Fisheye8k images with labels). To equip the Teacher Model with the capability to discern object distortions and size variations, we deploy a plethora of data augmentation techniques. These encompass methods like copy-paste, scaling, rotation, and perspective transformations, as depicted in Figure 4. Such techniques culminate in the formulation

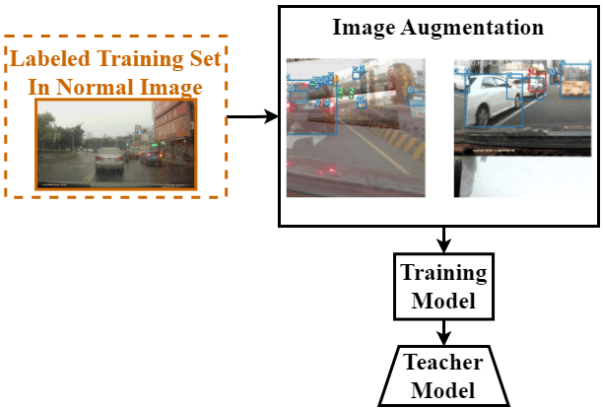

Figure 4. Training Teacher Model System Flowchart

of a robust teacher model, primed to address a broad spectrum of object representations. This teacher model is subsequently utilized on the test dataset for semi-automatic labeling.

We introduce a training strategy that melds knowledge distillation with semi-supervised learning anchored on pseudo-labels. The foundation is laid by crafting a resilient teacher model using training dataset supplemented by the aforementioned augmentation techniques. This teacher model then serves to generate pseudo-labels on previously unlabeled images. A subset of this pseudo-labeled data is meticulously refined to constitute a validation set for the upcoming student model. The training iteration for the student model is then set into motion. Progressively, as we record enhanced performance metrics on the validation set, it underscores the efficacy of the pseudo-labeled data in model optimization. The iterative dynamic of this process is visualized in Figure 5.

In the subsequent phase, we harness the pseudo-labeled data, instituting minor manual rectifications, before partitioning it into training and validation subsets. The teacher model is then engaged in knowledge distillation, facilitating knowledge transfer to a more compact student model. Following each training cycle of the student model, upon discerning a surge in validation performance, we curate fresh synthetic data to reinvigorate the training regimen. This semi-supervised modus operandi substantially slashes annotation overheads, a schematic of which can be appreciated in Figure 5.

3.4. Implementation Details

Our training strategy is meticulously crafted to navigate situations characterized by limited labeled data, while leveraging available models or data sources. Through this approach, our goal is to refine models that achieve a nuanced equilibrium between swift inference speeds and exceptional accuracy, ensuring timely completion of experiments.

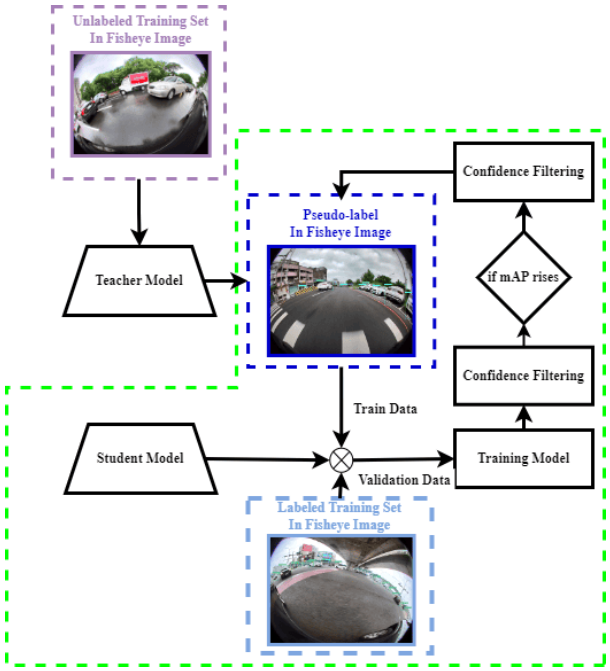

Figure 5. Training Student Model System Flowchart

A crucial step in the generation of pseudo-labeled data involves discerning the accuracy of annotations. To navigate this challenge, we employ the Confidence Filtering approach [24]. In this paradigm, the model's prediction confidence is of paramount significance. A high confidence from the model typically alludes to the veracity of its predictions, implying that they are dependable. In contrast, predictions with lower confidence necessitate rigorous filtering. Consequently, an adaptive threshold becomes indispensable – one that evolves in tandem with the model's refining accuracy. During the initial phases of model training, we intentionally keep the threshold on the lower side. This ensures we do not prematurely discard annotations related to smaller features, which might inherently have reduced confidence levels.

4. Results

4.1. Evaluation Model and Comparison with Existing Approaches

We have devised a comprehensive comparison chart delineating the efficacy disparity between employing Knowledge Distillation (KD) versus omitting it during the training phase. Notably, in the absence of KD, the recognition performance exhibits a noticeable shortfall, particularly concerning distant objects. Conversely, our proposed methodology facilitates the meticulous detection of all objects within the images, thereby attaining exceptional levels of recognition accuracy. For a more intricate elucidation, please consult Figure 6.

Table 1. FishEye8K testing table

Model	Version	Size	Precision	Recall	$mAP_{0.5}$	$mAP_{0.5-.95}$	F1-score	AP_S	AP_M	AP_L	Inference
YOLOv5	l6	1280	0.7929	0.4076	0.6139	0.4098	0.535	0.1299	0.434	0.6665	22.7
	x6	1280	0.8224	0.4313	0.6387	0.4268	0.5588	0.133	0.452	0.6925	43.9
YOLOR	W6	1280	0.7871	0.4718	0.6466	0.4442	0.5899	0.1325	0.4707	0.6901	16.4
	P6	1280	0.8019	0.4937	0.6632	0.4406	0.6111	0.1419	0.4805	0.7216	13.4
YOLOv7	D6	1280	0.7803	0.4111	0.3977	0.2633	0.5197	0.1261	0.4462	0.6777	26.4
	E6E	1280	0.8005	0.5252	0.5081	0.3265	0.6294	0.1684	0.5019	0.6927	29.8
YOLOv7	N	640	0.7917	0.4373	0.4235	0.2473	0.5453	0.1108	0.4438	0.6804	4.3
	X	640	0.7402	0.4888	0.4674	0.2919	0.5794	0.1332	0.4605	0.7212	6.7
YOLOv8	l	640	0.7835	0.3877	0.612	0.4012	0.5187	0.1038	0.4043	0.6577	8.5
	x	640	0.8418	0.3665	0.6146	0.4029	0.5106	0.0997	0.4147	0.7083	13.4

Table 2. Our testing table

Model	Version	Input Size	Precision	Recall	mAP0.5	mAP0.5-.95	f1-score
YOLOv7+**PRB(Ours)**	YOLOv7-E6E+PRB	640	0.6054	0.4012	0.4012	0.2501	0.2711
	YOLOv7-E6E+PRB	1280	0.7588	0.5212	0.5861	0.3304	0.6132
YOLOv6.3.0	YOLOv6-L	1280	0.7588	0.5212	0.5861	0.3304	0.6132
YOLOv6.3.0+PRB(Ours)	YOLOv6-L+PRB	1280	0.7731	0.4848	0.6264	0.3619	0.5959
YOLOv9	YOLOv9-E	640	0.7457	0.4720	0.5389	0.3118	0.5781
	YOLOv9-E	1280	0.7694	0.4829	0.5641	0.3287	0.5934
YOLOv9+PRB(Ours)	YOLOv9-E+PRB	640	0.7014	0.5257	0.56542	0.31965	0.6010
	YOLOv9-E+PRB	1280	0.7886	0.5203	0.3264	0.3795	0.6269

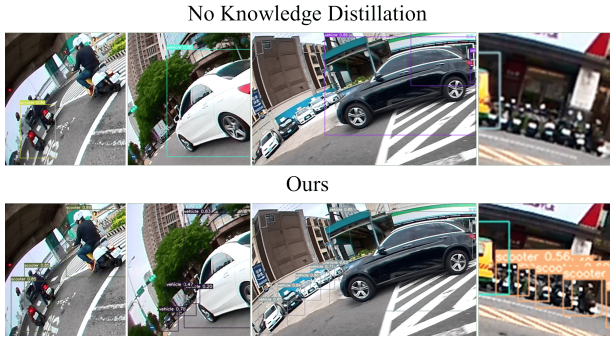

Figure 6. Performance Contrast: With vs Without Knowledge Distillation.

While the official FishEye8K benchmark results in Table 1 demonstrate impressive performance, our attempts to replicate these experiments fell short of fully reproducing the reported metrics. Nonetheless, we explored more recent iterations of the YOLO object detection family and introduced our proposed PRB (Probabilistic Region-Based) approach, which leverages knowledge distillation for data augmentation to enhance object localization capabilities.

As detailed in Table 2, incorporating our PRB method yielded improved results across the YOLOv6 3.0, YOLOv7, and YOLOv9 model variants when evaluated on the FishEye8K dataset. Specifically, our YOLOv6 3.0+PRB and YOLOv9+PRB models outperformed their counterparts in terms of mAP and F1-score metrics, demonstrating the efficacy of our region-based probabilistic framework for more accurate object detection on fisheye imagery. While our reproduced baselines underperformed the originally reported numbers, integrating the proposed PRB approach closed that gap and advanced the state-of-the-art on this challenging fisheye dataset.

Our YOLOv9+PRB model excels in detecting objects at road intersections, maintaining a decent accuracy in both day and night conditions. Despite this, occasional misclassifications and errors occur in certain images, prompting our commitment to further research for improvement. See Figure 7 for an example of our detection results.

5. Expected results

In this study, we participated in the CVPR competition and observed the dataset. We found a significant imbalance in the dataset, And anticipate that the method we proposed could effectively address this issue and accurately identify different road objects. Our recognition model tries to overcome the lighting differences between day and night. Even with fewer road images at night, our model may correctly identify objects. This achievement will help improve our ability to recognize road objects to the field of traffic moni-

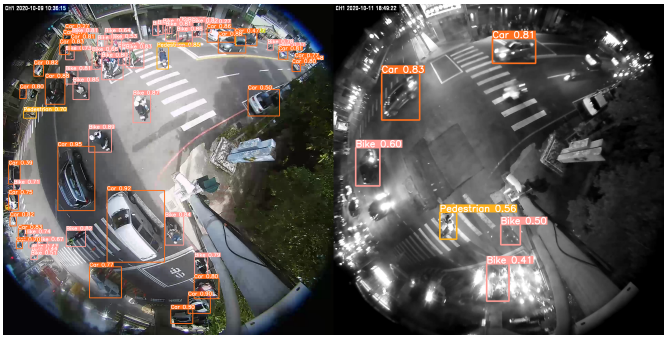

Figure 7. Fisheye Images Testing Result

toring systems. We hope that this research can contribute to the field of traffic monitoring systems, enhancing the ability to recognize road objects.

6. Conclusion

Through this report, we present our preliminary findings in fisheye image object recognition. We employ various performance-enhancing techniques, such as knowledge distillation (KD) and Parallel Residual Bi-Fusion (PRB) for image enhancement. Additionally, we utilize the YOLO series of object recognition models to conduct comparative analyses with the results obtained on the Fisheye8K dataset. Although our results do not match the performance reported in their experiments, to the best of our knowledge, other participants in the competition have also struggled to replicate their findings. Currently, we adopt a one-stage, lightweight strategy for rapid object recognition, which may impact the overall performance. In the future, we plan to explore the use of Transformers to train more extensive and accurate object recognition models, aiming to achieve state-of-the-art performance on fisheye imagery.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 1

[2] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. 2020. 3

[3] Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, and Yong-Sheng Chen. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE transactions on Image Processing*, 30:9099–9111, 2021. 2, 3

[4] Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, and Yong-Sheng Chen. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE Transactions on Image Processing*, 30:9099–9111, 2021. 2, 3

[5] Liuyuan Deng, Ming Yang, Yeqiang Qian, Chunxiang Wang, and Bing Wang. Cnn based semantic segmentation for urban traffic scenes using fisheye camera. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 231–236. IEEE, 2017. 1

[6] Gerardo Garcia-Gil and Juan M Ramirez. Fish-eye camera and image processing for commanding a solar tracker. *Heliyon*, 5(3), 2019. 1, 2

[7] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. pages 7029–7038, 2019. 3

[8] Myoung Gochoo, Zhenxue Wang, and Seoung Wug Cho. Fisheye8k and fisheye1keval: Large-scale datasets for road object detection in fisheye cameras. *arXiv preprint arXiv:2303.12745*, 2023. 1

[9] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, mar 2021. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015. 3

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015. 3

[12] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. SimPLE: Similar pseudo label exploitation for semi-supervised classification. jun 2021. 3

[13] Hyungtae Kim, Jaehoon Jung, and Joonki Paik. Fisheye lens camera based surveillance system for wide field of view monitoring. *Optik*, 127(14):5636–5646, 2016. 1, 2

[14] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu. Yolov6 v3. 0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023. 2

[15] 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. 2

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. pages 936–944, 2017. 3

[17] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. 2017. 3

[18] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 1

[19] Ajeet Ram Pathak, Manjusha Pandey, and Siddharth Rautaray. Application of deep learning for object detection. *Procedia computer science*, 132:1706–1717, 2018. 2

[20] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 2

[21] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1

[22] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 2

[23] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024. 2

[24] Huimin Wu, Xiaomeng Li, Yiqun Lin, and Kwang-Ting Cheng. Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023. 4

[25] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9308–9318, 2019. 1