Fusing Pseudo Labels with Weak Supervision for Dynamic Traffic Scenarios

Harshith Mohan Kumar[1]*, Sean Lawrence[2]†
[1]Department of Computer Science, PES University
[2]Intel Corporation

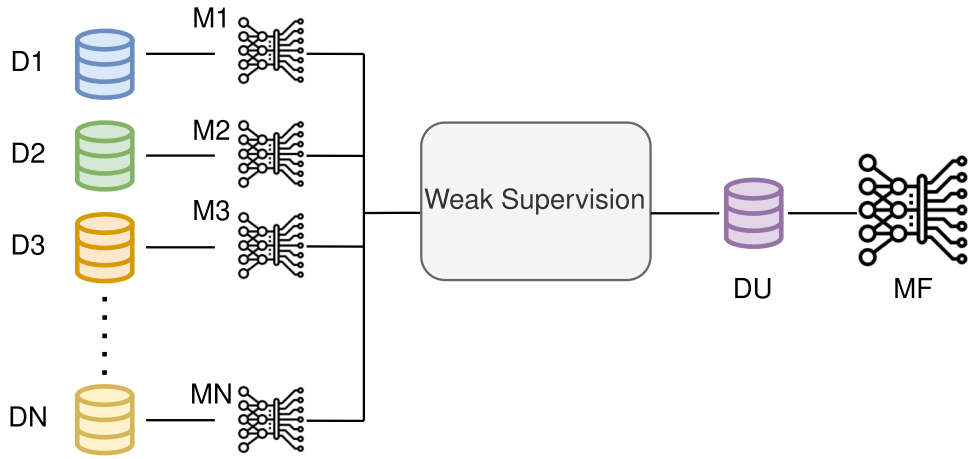

Figure 1: Proposed pipeline which merges pseudo labels from multiple models to form a unified dataset with weak supervision. The final model is trained on the unified label space.

Recent advancements in Advanced Driver Assistance Systems (ADAS) have witnessed remarkable progress and widespread adoption in the past decade. Historically, the automotive industry has been utilizing a fusion of various sensors, including lidar and radar [1]. However, drastic improvements in computer vision networks have enabled improved perception, enhanced decision-making capabilities and accurate prediction of impending collisions from camera inputs alone. Moreover, the application of vision based deep learning models has enabled ADAS to learn from large datasets, improving their ability to recognize and interpret complex driving scenarios [2].

While these systems have shown tremendous advancements, they face challenges in generalizing and adapting to various traffic conditions [3]. One of the main limitations of camera based models arises from the fact that large popular datasets used for training these models indirectly introduces bias of location, weather and traffic patterns. Additionally, the driving conditions found in these datasets revolve around well-maintained road infrastructure [4–7]. Changes in the environment, such as geographical locations,

*Work done while an intern at Intel Corporation, hiharshith18@gmail.com
†Corresponding author, sean.j.lawrence@intel.com

driving cultures and road infrastructure, can introduce distributional shifts [8]. These limitations ultimately degrade performance and reduce potential safety risks in situations that differ from the training data.

Several semi-supervised techniques such as pseudo labeling [9], contrastive learning [10,11], and noisy student [12] have demonstrated ways to learn a model from limited annotations. However, these techniques have strong dependence on the initial labeled data, sensitive to label noise and have difficulty in handling concept drift.

In this work we propose a weakly-supervised label unification pipeline with pseudo labels to train a singular object detection model from multiple datasets. Our pipeline architecture is illustrated in Figure 1. We initially fine-tune multiple homogeneous object detection models on each dataset. Then the final dataset D_U spanning the entire label space is populated using pseudo labels. Labels which fall under a certain threshold are manually checked to avoid propagating errors from the initial model. Finally we retrain a singular object detection model on the combined label space to produce a robust model M_F invariant to domain shifts.

Label Unification Pipeline. Following the work proposed in [13], we develop a label unification pipeline to combine N heterogeneous datasets, $D_1, D_2, ...D_N$ and corresponding their label spaces $L_1, L_2, ..., L_N$. The label spaces may consist of non-disjoint sets, $L_i \cap L_j \neq \emptyset$. These labels are merged and verified through a human operator. This process allows us to train a single model with the union of all label spaces $L_U = L_1 \cup L_2 ... \cup L_N$.

To produce L_U we initially fine-tune multiple detectors $M_1, M_2, ..., M_N$ on each dataset. The architecture of the model used remains the same. Each model populates the other $N-1$ datasets with pseudo labels above a certain threshold. We do not adopt the custom loss function proposed in [13], instead labels which fall under the threshold are flagged and passed to a verification process where a human annotator validates the true label. We use the Intel Geti tool for visual inspection.

Dataset. To demonstrate that our model can work in adverse road environments, we choose to gather road facing images from countries across Asia. For object detection we chose to work with the Indian Driving Dataset (IDD) [14] and Road Damage Dataset (RDD) [15–17]. Additionally we manually procure a dataset containing over 2600 ten second video clips recorded at the occurrence of a collision avoidance alert triggered by a Mobileye 8 Connect device.

Manual Alert Data. No public datasets consisting of Collision Avoidance Alert (CAS) alert metadata and scene frames are available for grading these alerts. To address this issue we develop a custom dataset that contains a diverse set of real-world driving scenes with various road types and collision scenarios. We generated over 2600 forward and pedestrian collision warnings across the city of Bangalore with varying lighting conditions.

We use the Mobileye 8 Connect system which is a popular choice for numerous personal and commercial vehicles globally. This device is an advanced automotive vision-based platform designed to enhance road safety.

Experiments. In this study we use two training procedures, one for training the object detection model and the other for weather classification. Our experiments were conducted on a system containing an Intel Core i5-9600K CPU paired with two Nvidia RTX 3060 GPUs with a total of 24GB of VRAM.

Model	Training	F1 Per Label									mAP (%)
		All	A	TS	M	R	P	C	Car	VF	
YoloX	S	0.42	0.49	0.19	0.45	0.38	0.24	0.69	.55	0.04	0.32
ATSS	S	0.61	0.73	0.53	0.62	0.55	0.50	0.79	0.72	0.24	0.54
SSD	S	0.40	0.40	0.01	0.48	0.37	0.18	0.56	0.53	0.06	0.28
Yolov8	S	0.60	0.75	0.73	0.69	0.68	0.54	0.80	0.79	0.32	0.65
Yolov8*	SSL	0.77	0.85	0.82	0.79	0.75	0.69	0.89	0.83	0.42	**0.78**

Table 1: Training results on the four different object detection networks. * Indicates the model trained on pseudo labeled dataset. Please see text for more details on the training sets and the baselines.

Class	Precision	Recall	mAP50	mAP50-95
all	0.887	0.677	0.808	0.613
A	0.927	0.785	0.886	0.738
TS	0.887	0.766	0.866	0.657
M	0.892	0.711	0.839	0.614
R	0.907	0.638	0.796	0.574
P	0.840	0.585	0.744	0.521
C	0.896	0.874	0.91	0.689
CA	0.887	0.772	0.866	0.708
VF	0.857	0.281	0.554	0.404

Table 2: Validation results of the chosen Yolov8 model (M_F) after training on pseudo labels.

We train and compare performance across four popular object detection networks, namely YoloX [18], Adaptive Sample Selection Training (ATSS) [19], Single Shot Multi-Box Detector (SSD) [20]. We used a batch size of 64, learning rate of 0.01, weight decay of 0.005 and trained our models for 50 epochs. The models are trained using the combination of Binary Cross Entropy, Bounding Box Loss and Dual Focal Loss.

We compare the class wise F1 scores and the mean average precision (mAP) on the training set across all four networks in Table 1. Of the four networks, the Yolov8 model performs significantly better than the rest. This model was then trained again on a label unified dataset using pseudo labels with manual verification on less confident predictions. We present the final validation set class-wise results in Table 2.

Conclusion. In this work we demonstrated the effectiveness of our weakly-supervised pseudo labeling pipeline in handling data distribution shifts. Our work can positively influence the accuracy of downstream ADAS tasks such as Collision Avoidance Alerts in areas with poor road infrastructure. In future work, we aim to demonstrate the capability of this pipeline for other computer vision tasks such as classification and segmentation.

References

[1] R. Okuda, Y. Kajiwara, and K. Terashima, "A survey of technical trend of adas and autonomous driving," in *Technical Papers of 2014 International Symposium on VLSI Design, Automation and Test*, 2014, pp. 1–4. 1

[2] S. Dabral, S. Kamath, V. Appia, M. Mody, B. Zhang, and U. Batur, "Trends in camera based automotive driver assistance systems (adas)," in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2014, pp. 1110–1115. 1

[3] F. M. Barbosa and F. S. Osório, "Camera-radar perception for autonomous vehicles and adas: Concepts, datasets and metrics," 2023. 1

[4] H. Schafer, E. Santana, A. Haden, and R. Biasini, "A commute in data: The comma2k19 dataset," 2018. 1

[5] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 11 618–11 628. 1

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[8] M. Hildebrand, A. Brown, S. Brown, and S. L. Waslander, "Assessing distribution shift in probabilistic object detection under adverse weather," *IEEE Access*, vol. 11, pp. 44 989–45 000, 2023. 2

[9] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning WREPL*, 07 2013. 2

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 2

[11] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020. 2

[12] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[13] X. Zhao, S. Schulter, G. Sharma, Y.-H. Tsai, M. Chandraker, and Y. Wu, "Object detection with a unified label space from multiple datasets," in *European Conference on Computer Vision (ECCV)*, 2020. 2

[14] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, jan 2019, pp. 1743–1751. 2

[15] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "Rdd2022: A multi-national image dataset for automatic road damage detection," 2022. 2

[16] ——, "Rdd2020: An annotated image dataset for automatic road damage detection using deep learning," in *Data in Brief*, vol. 36, 2021, p. 107133. 2

[17] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection and classification using deep neural networks with smartphone images: Road damage detection and classification," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, 06 2018. 2

[18] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," 2021. 3

[19] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 9756–9765. 3

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37. 3