Lightweight Object Detection Model with Optimal Transport Cost

Takeshi Uratsuka
Advanced Engineering Course
National Institute of Technology, Kurume College
Fukuoka, Japan
a4107tu@kurume.kosen-ac.jp

Kousuke Matsushima
Department of Control and Information Systems Engineering
National Institute of Technology, Kurume College
Fukuoka, Japan
matsushima@kurume-nct.ac.jp

***Abstract*—Municipal road maintenance is essential for safety, yet conventional inspection methods are costly and inefficient. Automated road damage detection has attracted attention, especially with the advancement of autonomous driving. However, most existing object detectors are trained under closed-set conditions, where all test classes are included during training. In realworld applications, unknown objects inevitably appear, and misclassifying them as known classes undermines system reliability. To address this, we study open-set object detection (OSOD) and propose improvements to the OpenDet-CWA framework and integrate efficient optimal transport–based distance metrics to improve feature compactness and separation in the embedding space. Specifically, we incorporate Max-Sliced Wasserstein, Markovian Sliced Wasserstein, and Random-Path Markovian Sliced Wasserstein distance into the class anchor alignment process. Furthermore, to enable real-time inference on resource-limited platforms, we employ RepViT and FastViT, a lightweight CNN architecture inspired by Vision Transformers, instead of conventional ResNet-50 backbone. Experimental evaluations on the VOC-COCO benchmark demonstrate that the proposed approach achieves robust open-set recognition while maintaining competitive accuracy in known-class detection. Additional experiments on a road damage dataset reveal that RepViT provides a favorable balance between accuracy and computational efficiency compared to FastViT, whose reduced representational capacity limits performance gains. Overall, the proposed method enhances open-set robustness and offers practical scalability for deployment in mobile devices and embedded road inspection systems.**

***Index Terms*—RepViT, FastViT, Optimal Transport Cost, Supervised Learning, Road Damage Detection**

I. Introduction

Ensuring the structural integrity and operational safety of road infrastructure is a critical issue in modern transportation systems. Municipal roads, which constitute roughly 80% of total road length used for daily travel [1], are predominantly managed by local governments. However, manual inspection remains the primary means of assessing road surface conditions, and this process is inherently time-consuming, labor-intensive, and costly. Consequently, inspection frequency and spatial coverage are often insufficient to prevent the progression of minor defects into serious pavement failures. Surface degradation such as cracking, rutting, or potholes can disrupt drainage, cause delamination between layers, and lead to foundation settlement—ultimately shortening the service life of roadways.

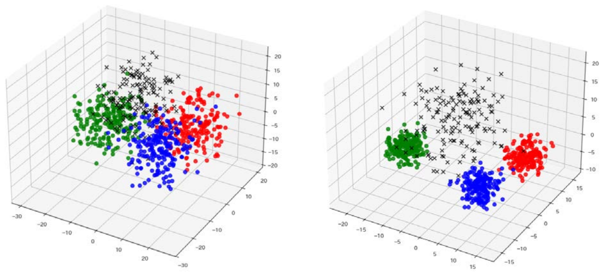
To mitigate these issues, numerous studies have explored automated road damage detection using image-based and sensor-based approaches. The introduction of deep learning has accelerated this trend, enabling significant advances in visual recognition accuracy. In parallel, the widespread deployment of autonomous vehicles and mobile sensing technologies has created a growing demand for reliable and real-time road condition assessment. Traditionally, object detection models in this field have been developed under closed-set assumptions, where training and test datasets contain identical categories. Collaborative projects between research institutions and municipalities have resulted in large-scale annotated datasets consisting of thousands of road surface images captured via smartphones. Utilizing these datasets, deep learning–based detectors—particularly one-stage architectures such as the YOLO family—have demonstrated impressive accuracy and real-time performance in recognizing various damage types [2], [3] .

However, real-world environments are inherently open-set, meaning that models inevitably encounter unknown or unseen damage categories not present in training data. Such conditions severely degrade the reliability of conventional detectors, which tend to misclassify unfamiliar damage patterns as known classes. This misclassification poses significant safety risks for downstream applications such as autonomous driving or infrastructure monitoring. To address this challenge, research attention has shifted toward Open-Set Object Detection (OSOD), which explicitly accounts for the presence of unknown samples during inference [4], [5].

Recent studies have proposed several frameworks to enhance open-set robustness. For instance, the Open-World Detection Transformer (OW-DETR) integrates novelty scoring and objectness estimation within a Transformer-based pipeline, enabling the recognition of unseen objects [6]. Meanwhile, the Unknown-Classified Open-World Object Detection (UC-OWOD) framework introduces multi-cluster grouping for unknown instances and an improved evaluation protocol for unknown-class detection [7].

Building upon these advances, this study adopts the

979-8-3315-7896-1/26/$31.00 ©2026 IEEE



(a) (b)

Fig. 1. A 3D diagram visualizing the embedded space. Round dots represent objects belonging to known classes, while cross dots represent objects belonging to unknown classes. Here, three known classes are represented in red, blue, and green, respectively. The goal is to compact the clusters of objects represented by round points belonging to the same known class and reduce the size of the known class clusters. In other words, it is to bring Figure (a) closer to the state shown in Figure (b). By doing so, the area represented by crosses for unknown class can be expanded, thereby improving the accuracy of unknown class estimation.

OpenDet-CWA (OD-CWA) framework [8] and introduces multiple architectural and algorithmic refinements aimed at improving both accuracy and computational efficiency. Specifically, we replace the conventional optimal transport (OT) distance in the loss function with a more computationally efficient alternative. Our objective is to enhance feature compactness among known categories while expanding the representation margin that surrounds unknown regions in the embedding space. This strategy promotes clearer class separation and improved unknown-class detection capability.

Building upon our previous work [9], which introduced the Markovian and random-path sliced Wasserstein distances for efficient optimal transport in open-set detection, this study further integrates these cost functions with lightweight CNN and hybrid Transformer backbones (RepViT, FastViT) to achieve real-time inference on mobile platforms. Experimental results demonstrate that the proposed approach achieves consistent accuracy gains and a substantial reduction in computation time, contributing to the development of a more practical and robust mobile road-damage detection system.

II. Conventional Method

A. OpenDet-CWA(OD-CWA)

The OpenDet-CWA (OD-CWA) is an enhanced Open-Set Object Detection (OSOD) framework that mitigates misclassification of unknown objects by separating high-density latent regions (knowns) from low-density ones (unknowns). As an example, Fig. 1 visualizes the latent space features of the three known classes (colored circles) and the unknown class (crosses). Built upon Open-Det with a Faster R-CNN backbone (e.g., ResNet-50 + RPN), OD-CWA integrates four core modules: Spectral Normalization (SN) [10], Contrastive Feature Learner (CFL) [11], Class Wasserstein Anchor (CWA), and Unknown Probability Learner (UPL). SN stabilizes the final linear layer by normalizing weight spectra, preserving distance awareness between training and test samples. CFL enforces intra-class compactness and inter-class separation via instance-level contrastive learning, thus concentrating known features and isolating unknowns. CWA introduces a Wasserstein-based loss to align logits with class anchors, improving boundary precision and feature compactness. Finally, UPL estimates an explicit “unknown probability” per instance, using prediction uncertainty as a threshold to distinguish low-density unknown regions. Collectively, these components refine feature embeddings and enhance discrimination between known and unseen object classes.

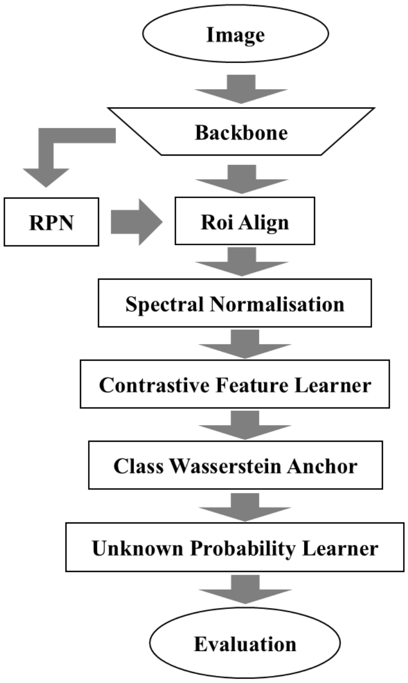


Fig. 2. Flow chart of OD-CWA. The CFL components utilises proposal features encoded into low-dimensional embeddings using the Contrastive Head (CH) optimised using Instance Contrastive Loss. The weights of the linear output layer are passed through a SN step that maintain distance awareness property. Then UPL component utilises the cosine distances between embeddings and spectral normalised weights to learn the probabilities for both known classes (C_K) and the unknown class (C_U). The class Wasserstein anchor part aids both CFL & UPL to increase the compactness in the clusters by finding the optimal transport plan.

B. Optimal Transport Cost

The optimal transport cost represents the minimum effort required to transform one probability distribution into another, reflecting not just shape similarity but the geometric movement of probability mass. It serves as a key metric for comparing distributions, often used in generative model evaluation and as a loss function. In this work, the Wasserstein distance is adopted as the optimal transport cost, as it measures the discrepancy between predicted and anchor distributions [13]. However, because its computational cost grows with dimensionality, we employ a slicing-based approximation for efficiency. The slicing method projects high-dimensional distributions onto one-dimensional directions, enabling fast computation of transport costs. We propose three efficient variants: the max-sliced(Max-SW) Wasserstein distance, which focuses on the most discriminative projection, and the Markovian sliced Wasserstein distance(MSW), which exploits Markovian dependencies to avoid redundant projections. Additionally, Random-Path Markovian sliced Wasserstein distance(RP-MSW) which employs the Random-Path strategy further enhances more effective recognition of the geometric structure of distributions.

1) Max-Sliced Wasserstein distance: The Max-sliced Wasserstein distance (Max-SW) [14] is an efficient extension of the sliced Wasserstein distance (SW) designed to reduce the computational cost of the traditional Wasserstein metric. It works by projecting two high-dimensional probability distributions onto multiple one-dimensional directions and computing the Wasserstein distance for each projection. The maximum of these distances is then taken as the Max-SW, as it most effectively captures the distinguishing features between distributions. By optimizing projection directions toward this maximum, Max-SW highlights the most significant differences while preserving essential distributional information.

$$Max\text{-}SW_p(\alpha,\beta)= \underset{\theta\in\mathbb{S}^{d-1}}{max} W_p(\theta_\sharp\alpha,\theta_\sharp\beta) \tag{1}$$

α and β are input distributions, and projection direction θ is sampled from a unit hypersphere of $\mathbb{S}^{d-1}$ dimensions. Normally, a one-dimensional Wasserstein distance with p=1 is calculated. Since the projection direction component is optimized in the direction of the greatest distance, maximizing the Wasserstein distance maximizes the distance between distributions and allows us to capture the differences between distributions.

2) Markovian Sliced Wasserstein distance: The Markovian Sliced Wasserstein Distance (MSW) [15] addresses the limitations of Max-SW by introducing sequential dependencies among projection directions. Instead of finding a single optimal projection, MSW imposes a first-order Markov structure, where each projection direction depends on the previous one. This balances the randomness of sampling with the optimization of Max-SW, efficiently identifying informative projections with fewer samples.

$$MSW^p_{p,T}(\alpha,\beta)=\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p(\theta_t\sharp\alpha,\theta_t\sharp\beta)\right] \tag{2}$$

MSW computes the average p-Wasserstein distance over T steps by sampling projection directions from a Markovian distribution σ, typically using Markov Chain Monte Carlo (MCMC) for efficient sampling.

3) Random-Path: In Max-SW, finding the optimal projection direction requires costly optimization. To address this, the Random Path (RP) method [16] offers a more efficient alternative. It randomly samples one vector from each of the two distributions, computes their difference $Z = X - Y$, and normalizes it to obtain the Random-Path Projecting Direction (RPD). This direction approximates the divergence between distributions without optimization, achieving faster and more stable computation compared to Max-SW.

III. Proposed Method

A. Overview

To achieve real-time open-set object detection on resource-limited devices while maintaining high recognition robustness, we propose an enhanced OD-CWA framework that integrates efficient optimal transport (OT) losses with lightweight hybrid CNN–Transformer backbones. Unlike conventional OD-CWA, which relies on a heavy ResNet-50 architecture, our method introduces adaptive OT-based feature alignment within compact models such as RepViT and FastViT. This design aims to (1) preserve inter-class separability and intra-class compactness in the latent space, and (2) minimize computational overhead by applying OT regularization selectively to high-level embeddings rather than all feature maps.

B. Lightweight Backbone Adaptation

1) FastViT Integration: FastViT [17] is a hybrid vision transformer that balances latency and accuracy by combining transformer-style token mixing with efficient convolutional operations. It employs three key design strategies: the RepMixer block, which replaces MetaFormer skip connections with a reparameterized depthwise convolution at inference to reduce latency; linear train-time overparameterization, which temporarily adds extra branches to enhance feature capacity and is later merged into a single path; and large depthwise kernels to expand the receptive field and improve robustness to out-of-distribution samples. With this four-stage architecture, FastViT achieves 83.9% Top-1 ImageNet accuracy while running up to 1.9 × faster than ConvNeXt on mobile hardware, maintaining strong performance across object detection and segmentation tasks.

2) RepViT Integration: While OD-CWA originally used ResNet-50, replacing it with RepViT [18] results in a lighter and faster model suitable for mobile deployment. RepViT is a lightweight CNN inspired by Vision Transformer design principles, implemented within a MetaFormer framework using fully re-parameterized convolutions. Its core block separates token and channel mixing through depthwise re-parameterization, eliminating skip-connection overhead during inference and reducing computational cost. At the architectural level, RepViT employs a simplified stem, deeper downsampling with RepViT and FFN blocks, and a minimal global-average-pooling classifier to further improve latency. It favors efficient 3×3 convolutions, a small expansion ratio, increased channel width, and selectively placed SE layers for a balance of speed and accuracy. RepViT achieves strong accuracy–efficiency trade-offs, surpassing prior lightweight CNNs and ViTs, exceeding 80% Top-1 accuracy on ImageNet with only 1.0 ms latency, and demonstrating robust performance in detection and segmentation tasks.

C. *Summary of Advantages*

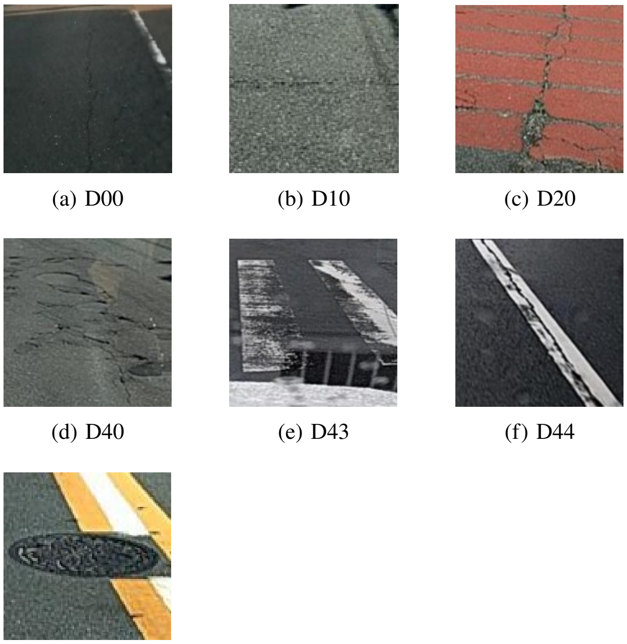
The proposed modifications provide the following benefits:

- Efficiency: Selective OT computation and lightweight backbones drastically reduce training and inference costs.
- Robustness: Wasserstein-based alignment improves discrimination of unknown samples by expanding low-density latent regions.
- Scalability: The unified OT–RepViT/FastViT framework maintains stable performance across different computational budgets and datasets.

IV. Experiment

A. *Experimental Setup*

In this study, we utilized the VOC-COCO dataset, a combination of Pascal VOC [19] (2007, 2012) and MS COCO [20] (2017). In the datasets, VOC-COCO includes 20 known classes, while its open-set variants—VOC-COCO-20, VOC-COCO-40, and VOC-COCO-60—contain 20 known and 20, 40, and 60 unknown classes, respectively, with approximately equal numbers of images per class. Thus, VOC-COCO serves as a closed-set dataset, whereas the others simulate open-set conditions with varying numbers of unknown categories. Additionally, we employed a proprietary Road Damage Dataset (RDD) consisting of seven road damage types, such as cracks and potholes. Examples are shown in Figure 3. The RDD is an adapted version of the Road Damage Dataset [21], originally designed for object detection; each bounding box was cropped for classification use. To evaluate open-set performance, we adopted Wilderness Impact (WI) [22], which measures the misclassification rate of unknowns as knowns, and Absolute Open-Set Error (AOSE) [23], which quantifies the total number of such misclassifications. In addition, mean Average Precision for known classes (mAP_K) and Average Precision for unknowns (AP_U) were used to assess classification accuracy.


(a) D00

(b) D10

(c) D20

(d) D40

(e) D43

(f) D44

(g) D50

Fig. 3. We used road damage dataset(RDD) including potholes, cracks, fissures and so on. Please note that this dataset is a proprietary dataset.

TABLE I System Specifications.

OS	Windows 11 Pro
CPU	Intel Core i7-14700 @ 2.10 GHz
RAM	32 GB
GPU	NVIDIA GeForce RTX 4080 Super with 16 GB VRAM

Lower WI and AOSE values indicate better discrimination, while higher mAP_K and AP_U reflect improved accuracy. The mAP_K metric is defined as $mAP_K = \frac{1}{N}\sum_{i=1}^{N} AP_i$, where N denotes the number of known classes. The experimental hardware and software configurations are summarized in Table I.

B. *Main Results*

Table II presents a quantitative comparison of various methods on the VOC-COCO benchmark using three different backbones. When using ResNet-50, Max-SW and RP-MSW achieve the highest mAP values, indicating strong performance in known-class detection, while RP-MSW slightly reduces AOSE and WI compared to other approaches, suggesting better robustness against open-set errors. OD-CWA attains competitive mAP but exhibits larger AOSE, implying a higher tendency to misclassify unknown samples as known ones. In contrast, RP-MSW maintains comparable accuracy with substantially reduced open-set misclassification, confirming the effectiveness of the reconstruction-based prior in stabilizing the decision boundary.

TABLE II Comparisons with various methods on VOC-COCO.

(a) Using ResNet-50 as backbone.

Dataset	VOC-COCO	VOC-COCO-20				VOC-COCO-40				VOC-COCO-60				-
Method	mAP	WI	AOSE	AP_k	AP_U	WI	AOSE	AP_k	AP_U	WI	AOSE	AP_k	AP_U	Training time [s/itr]
OD-CWA	76.83	10.7	12375	56.16	13.88	13.42	19562	53.03	10.4	12.23	26479	53.48	4.77	6.58
Max-SW	77.18	10.44	11991	56.14	15.13	12.88	18972	53.03	10.77	11.77	25947	53.57	4.68	16.48
MSW	76.56	10.93	12717	55.89	15.52	13.79	20233	52.86	11.05	12.41	27060	53.34	4.86	35.93
RP-MSW	77.72	10.91	11705	56.66	15.11	13.81	18446	53.54	10.87	12.21	24934	53.88	4.83	37.52

(b) Using RepViT as backbone.

Dataset	VOC-COCO	VOC-COCO-20				VOC-COCO-40				VOC-COCO-60				-
Method	mAP	WI	AOSE	AP_k	AP_U	WI	AOSE	AP_k	AP_U	WI	AOSE	AP_k	AP_U	Training time [s/itr]
OD-CWA	26.09	7.73	13133	17.64	8.89	8.65	18280	16.71	6.67	6.78	22393	17.04	3.46	0.11
Max-SW	26.33	7.07	12164	17.6	7.74	8.04	17597	16.66	5.95	6.56	22021	17.02	3.2	1.48
MSW	27.46	7.72	13091	17.29	8.16	8.73	18419	16.24	6.09	6.71	22215	16.69	3.08	4.31
RP-MSW	25.07	7.36	12774	16.53	7.46	8.23	17746	15.88	5.76	6.37	20920	16.05	3.12	4.45

(c) Using FastViT as backbone.

Dataset	VOC-COCO	VOC-COCO-20				VOC-COCO-40				VOC-COCO-60				-
Method	mAP	WI	AOSE	AP_k	AP_U	WI	AOSE	AP_k	AP_U	WI	AOSE	AP_k	AP_U	Training time [s/itr]
OD-CWA	6.9	6.11	6432	2.98	3.04	7.75	10457	2.71	2.6	6.65	12309	2.68	1.27	0.14
Max-SW	16.64	5.34	7703	8.79	4.09	6.51	11192	8.23	3.48	4.87	12264	8.33	1.63	1.51
MSW	13.34	4.95	5894	6.61	3.16	5.84	8244	6.28	2.89	4.42	9224	6.36	1.46	4.16
RP-MSW	17.95	5.6	7974	8.99	3.72	6.83	11715	8.36	3.25	5.13	12729	8.5	1.62	4.44

TABLE III Comparisons with various methods on RDD. In the RDD experiments, lightweight backbone was adopted to facilitate efficient real-time processing. Although seven known classes were used for training in RDD, they were not treated as unknown classes during testing. Therefore, the WI, AOSE, and AP_U evaluation metrics are not reported.

Dataset	RDD			
Method	RepViT		FastViT	
	mAP	Training time[s/itr]	mAP	Training time[s/itr]
OD-CWA	28.18	0.10	26.12	0.12
Max-SW	29.59	1.43	27.25	1.39
MSW	27.97	3.92	23.92	4.07
RP-MSW	27.80	4.48	18.89	4.44

With RepViT as the backbone, the overall performance decreases compared to ResNet-50, reflecting the relatively weaker feature separability of lightweight transformer architectures. Nevertheless, MSW achieves the highest mAP, demonstrating the benefit of Wasserstein-based feature regularization for open-set generalization. RP-MSW, while showing slightly lower mAP, achieves the smallest AOSE, indicating its superior ability to suppress false recognitions of unknown objects.

Using FastViT, all methods exhibit further degradation in both known and unknown detection performance, which is expected given its compact architecture. Even in this setting, RP-MSW consistently yields the lowest AOSE and WI, highlighting its robustness under limited representational capacity. OD-CWA achieves relatively higher AP_k but suffers from a sharp drop in AP_u, confirming its bias toward known-class discrimination.

Overall, across all three backbones, RP-MSW consistently provides the best balance between maintaining known-class accuracy and minimizing open-set recognition errors, demonstrating its general applicability and stability in various network architectures.

Table III summarizes the experimental results on the RDD dataset using lightweight backbones to emphasize computational efficiency. Among the evaluated methods, Max-SW achieved the highest mAP across both RepViT and FastViT, demonstrating its effectiveness in preserving discriminative power even under reduced model capacity. OD-CWA, while exhibiting the fastest training speed with 0.10 s/iteration for RepViT and 0.12 s/iteration for FastViT, showed relatively lower accuracy, indicating a trade-off between computational efficiency and recognition performance.

MSW and RP-MSW required substantially longer training times due to the additional computation of Wasserstein-based losses and reconstruction priors, respectively. In particular, RP-MSW exhibited the highest training cost (4.48 s/iteration with RepViT and 4.44 s/iteration with FastViT) but did not yield corresponding improvements in mAP, suggesting that the reconstruction-based prior was less effective under the lightweight backbone constraint. This result implies that while RP-MSW provides strong robustness in open-set scenarios, its benefits are diminished when model capacity and feature expressiveness are limited.

Overall, the RDD experiments highlight that in resource-constrained settings, simpler Wasserstein-based methods such as Max-SW can achieve a favorable balance between accuracy and efficiency, whereas the heavier reconstruction-based variant may not provide proportional gains in detection performance.

V. Discussion

The results in Tables II and III highlight the varying effectiveness of optimal transport–based methods across different backbone architectures and datasets. On the VOC-COCO benchmarks, the proposed RP-MSW consistently achieves performance comparable to or better than baseline methods, indicating robust open-set recognition by promoting compact known-class representations while maintaining separation from unknowns. However, on the lightweight RepViT and FastViT backbones used in the RDD experiments, the advantages of RP-MSW diminish due to its increased computational overhead and sensitivity to reduced feature capacity. In these efficiency-constrained settings, Max-SW provides a more practical balance, achieving strong accuracy with substantially lower training cost. Overall, RP-MSW is effective with standard-capacity backbones such as ResNet-50, but applying it to real-time or resource-limited systems requires further refinement or simplification.

VI. Conclusion

In this study, we improved the computational cost by introducing lightweight backbones such as RepViT and FastViT in open-set object detection across multiple datasets. Particularly on complex benchmarks like VOC-COCO, RP-MSW demonstrated consistently high performance, confirming its effectiveness in improving separability between known and unknown classes in feature space. However, experiments on RDD revealed that combining lightweight backbones with RP-MSW significantly increases computational cost. This indicates that RP-MSW's advantages depend on the backbone's representational capacity, and lightweight models tend to disrupt the balance between effectiveness and efficiency. Therefore, a future challenge is to explore more efficient methods that maintain reconstruction benefits while reducing computational load. These improvements are expected to enable the stable application of open-set object detection models even in real-time operational environments such as mobile devices and embedded systems.

VII. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 25K15171.

References

[1] Japanese Road Statistics Annual Report, Ministry of Land, Infrastructure, Transport and Tourism, 2021.

[2] Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiyama, Hiroshi Omata, ”Road Damage Detection Using Deep Neural Networks with Images Captured Through a Smartphone,” arXiv: 1801.09454, Feb., 2018.

[3] Wei Wang, Xiaoru Yu, Bin Jing, Ziqi Tang, Wei Zhang, Shengyu Wang, Yao Xiao, Shu Li, Liping Yang, ”YOLO-RD: A Road Damage Detection Method for Effective Pavement Maintenance,” Sensors, vol. 25, no. 5, Article 1442, 2025, doi: 10.3390/s25051442.

[4] J. Han, Y. Ren, J. Ding, X. Pan, K. Yan, G. Xia, ”Expanding Low-Density Latent Regions for Open-Set Object Detection,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9591-9600, 2022.

[5] K J Joseph, Salman Khan, Fahad Shahbaz Khan, Vineeth N Balasubramanian., ”Towards open world object detection,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5830-5840, 2021.

[6] Akshita Gupta, Sanath Narayan, K J Joseph, Salman Khan, Fahad Shahbaz Khan, Mubarak Shah, ”OW-DETR: Open-world Detection Transformer,” arXiv: 2112.01513, Apr., 2022.

[7] Zhiheng Wu, Yue Lu, Xingyu Chen, Zhengxing Wu, Liwen Kang, Junzhi Yu, ”UC-OWOD: Unknown-Classified Open World Object Detection,” arXiv: 2207.11455, Jul., 2022.

[8] P. Mallick, F. Dayoub, J. Sherrah, ”Wasserstein Distance-based Expansion of Low-Density Latent Regions for Unknown Class Detection,” Australian Institute of Machine Learning North Terrace, Adelaide SA 5000, Jan., 2024.

[9] Takeshi Uratsuka, Masahiro Iwahashi, Ryosuke Harakawa, Kousuke Matsushima, ”Unknown-Class Road Damage Classification with Optimal Transport,” THE 17th INTERNATIONAL CONFERENCE ONINFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING (ICITEE), pp. 348-353, Oct., 2025.

[10] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. ”Simple and principled uncertainty estimation with deterministic deep learning via distance awareness,” In Advances in Neural Information Processing Systems, pp. 7498–7512, 2020.

[11] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, ”Supervised Contrastive Learning,” in Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 18661–18673, Dec., 2020.

[12] K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[13] D. Miller, N. Sünderhauf, M. Milford, and F. Dayoub, ”Class Anchor Clustering: A Loss for Distance-based Open Set Recognition,” IEEE Winter Conference Applications of Computer Vision, pp. 3569–3577, 2021.

[14] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, Alexander Schwing, ”Max-Sliced Wasserstein Distance and its use for GANs,” arXiv: 1904.05877, Apr., 2019.

[15] Khai Nguyen, Tongzheng Ren, Nhat Ho, ”Markovian Sliced Wasserstein Distances: Beyond Independent Projections,” International Conference On Machine Learning, pp 39812 - 39841, Dec., 2023.

[16] Khai Nguyen, Shujian Zhang, Tam Le, Nhat Ho, ”Sliced Wasserstein with Random-Path Projecting Directions, ” International Conference On Neural Information Processing Systems, pp 37879-37899, Jul., 2024.

[17] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, Anurag Ranjan, ”FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization, ” arXiv: 2303.14189, Aug., 2023.

[18] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, Guiguang Ding, ”RepViT: Revisiting Mobile CNN From ViT Perspective, ” arXiv: 2307.09283, Mar., 2024.

[19] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision (IJCV), 2010; 88: 303–338.

[20] 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 ECCV 2014. Springer, Cham. pp. 740–755.

[21] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, ”Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images, ” Computer-Aided Civil and Infrastructure Engineering, 33, 12, pp.1127-1141, Jun., 2018.

[22] A. R. Dhamija, M. Günther, J. Ventura and T. E. Boult, ”The Overlooked Elephant of Object Detection: Open Set,” 2020 IEEE Winter Conference on Applications of Computer Vision, pp. 1010-1019, 2020.

[23] D. Miller, N. Sünderhauf, M. Milford and F. Dayoub, ”Uncertainty for Identifying Open-Set Errors in Visual Object Detection,” IEEE Robotics and Automation Letters, vol. 7, no. 1, pp. 215-222, Jan,. 2022.