# Field study toward anomaly road damage detection with drive recorder

Masato Tsuchiya [1], Ken Miyamoto [1], Takashi Ota [1], and Yasushi Sugama[1]

[1] *Mitsubishi Electric Information Technology R & D Center*
*miyamoto.ken@bc.mitsubishielectric.co.jp*
*Ota.Takashi@dx.MitsubishiElectric.co.jp*
*sugama.yasushi@bp.mitsubishielectric.co.jp*

## ABSTRACT

As one of the ways to reduce road maintenance costs, road damage detection with a mobile camera is gaining attention. Most of conventional damage detection use supervised learning, nevertheless three practical drawbacks exist. Firstly, supervised learning requires a high manual cost to collect annotated data for training. Secondly, some damages are rarely observed, resulting in imbalanced data and difficulty in training an efficient model for all damage categories. Additionally, annotators may not identify such rare damages correctly. Thirdly, supervised learning cannot detect unknown categories of damages, though unknown categories are often found in a practical scene. To overcome these three drawbacks, we propose an ensemble model that combines anomaly detection and supervised damage detection. Anomaly detection can detect previously unknown and rare types of damage, while supervised damage detection ensures damages frequently observed on roads. Two different models cover wider categories of road damages. Our ensemble model is expected to achieve higher accuracy and lower manual cost.

## 1. INTRODUCTION

Automated health monitoring is expected to reduce maintenance costs in civil infrastructure industry. Examples of the monitoring are roads, bridges, and power insulators. Machine learning is a typical approach to detect a damage on above-mentioned civil infrastructure, such as roads (Maeda, Sekimoto, Seto, Kashiyama, & Omata, 2018), bridges (Bukhsh, Jansen, & Saeed, 2021), and power line insulators (Tao et al., 2018). These conventional works mainly use one of supervised learning algorithms. There are three negative aspects to apply supervised learning on an application of civil infrastructure. Firstly, damages of civil infrastructures are infrequent, making it expensive to collect these data for model training and evaluation. Secondly, a special equipment (e.g., drone) is required to acquire data because maintenance associates can't reach overall areas on a large bridge or dam walls by their foots. Occasionally, further consideration is forced to apply supervised learning due to difference between a commoditized camera and a sensor on a special equipment. Finally, supervised learning can't make a model to detect unknown categories, though unknown damages may exist practically. For resolving the three issues, we propose an ensemble model of supervised damage detection and anomaly detection.

The contributions of this paper are summarized as follows:

1. We propose an ensemble road damage detection to find both known categories of damages and unknown (i.e., versatile minor) phenomena. That may make effort to capture versatile damages on foreground objects (e.g., road markers).
2. We create a dataset taken by a drive recorder, which assumes collaborative data collection from vehicles.
3. This paper shows challenges, when applying anomaly detection to road maintenance problem. The challenges have not been shown in other papers.

## 2. RELATED WORK

For detecting an object from a single image, Convolutional Neural Network (CNN) approaches have been developed, such as YOLO-series (Wang, Yeh, & Liao, 2021; Ge, Liu, Wang, Li, & Sun, 2021; Wang, Bochkovskiy, & Liao, 2022). A typical feature of the above CNNs can work on little computational resource. Hence, the light CNNs are implemented even on an embedded machine. Some of Road Damage Detection (RDD) have been developed on the light CNNs, because road industry expects to detect a road damage on an embedded machine of vehicles (Jo & Ryu, 2015; Nienaber, Booysen, & Kroon, 2015). However, the light CNNs can't detect diverse minor damages often observed on roads, because CNNs are modeled under known categories. Therefore,

we attempted to combine supervised RDD and anomaly detection.

Anomaly detection can be classified into three types: reconstruction-based, similarity-based, and flow-based. In reconstruction-based approaches, normal states are modeled through unsupervised learning, allowing the detection of anomalies even with limited anomaly samples. Examples of reconstruction-based algorithms include Autoencoders, Variational Autoencoders, and Generative Adversarial Networks (Akcay, Atapour-Abarghouei, & Breckon, 2019). Similarity-based approaches such as PaDiM (Defard, Setkov, Loesch, & Audigier, 2021) and CFA (Lee, Lee, & Song, 2022) map input vectors into a lower dimensional embedding space. Anomalous objects are identified by their deviation from normal states in this space. Flow-based approaches such as CSFlow (Rudolph, Wehrbein, Rosenhahn, & Wandt, 2022) transform normal states into a lower dimensional distribution, which is formulated as an arbitrary probabilistic density function through multiple invertible transformations. In this context, anomalous objects can be detected on low probability points.

## 3. Proposed Hybrid framework

### 3.1. Ensemble of object detection and anomaly detection

We present our ensemble method for detecting both known categories of damages and unknown phenomena. The pipeline of our method is depicted in Figure 1. We utilized Test Time Augmentation (TTA) and Weighted Boxes Fusion (WBF) (Solovyev, Wang, & Gabruseva, 2021) in our ensemble method during inference. Combination of TTA and WBF is frequently used to improve accuracy of object detection task including supervised RDD. Details of TTA and WBF are shown in (Zhao, Zhang, & Zhao, 2023; Hegde et al., 2020). We selected the state-of-the-art object detection models (YOLOX, YOLOv7 and YOLOv8) and anomaly detection models (CFA, PaDiM and CSFlow). These models are carefully selected to ensure diverse types of algorithms.

In order to fuse supervised RDD and anomaly detection, outcomes of both models will be merged. There is one obstacle to merge both outcomes. An outcome of anomaly detection model is heat map, which indicates deviation from normal states. On the other hand, an outcome of supervised RDD is bounding boxes. For fusing different types of outcomes, the heat map is turned into bounding boxes before fusion. First, the heat map is transformed to a segmentation mask by a predefined threshold. Second, boundaries of the segmentation mask fit rectangles. The above two steps generate bounding boxes of anomaly detection.

### 3.2. Preprocessing and Pretraining

**Preprocessing**: Images captured by a drive recorder contain a lot of roadside objects(e.g., glasses, shops or rails). To reduce influence of the roadside objects, all images are cropped to focus on road surfaces, as shown in Figure 2. Each image is cropped 20 to 60% vertically, then the cropped image is resized to adjust an original image size. Next, the image is cropped into a trapezoidal shape to remove areas outside of the road. Although some works apply road segmentation (Xu, Xiong, & Bhattacharyya, 2022), we don't use the segmentation because of possibility to remove a road region, unexpectedly.

When an anchor-based object detector is selected as supervised RDD, anchor boxes are optimized by classical differential evolution algorithm before training. Otherwise, we don't use any optimizations.

**Pretraining**: We use pre-trained weights for both supervised RDD and anomaly detection. Supervised RDD pre-trained MSCOCO that targets object detection. Anomaly detection pre-trained ImageNet that aims image classification task.

## 4. Experiments

### 4.1. Datasets and Metrics

**Datasets**: We created two datasets for our evaluation. The first dataset is a set of images captured by a drive recorder. We installed the drive recorder on a windshield of a vehicle. Standard HD (1280 * 720) is resolution of a camera on the drive recorder. We collected 11,536 images (i.e., 2549 anomaly images and 8987 normal images) on National Highway 9 in Japan. After the image collection, we annotated bounding boxes on damages, when the damages are clearly present. The annotated images are split into training and test sets at a ratio of 9:1.

The second dataset is synthetically created. Road damages are extracted from the first dataset, as shown in Figure.3. The extracted damages are overlaid onto undamaged regions. As an overlaying method, we apply Paste and Learn (Dwibedi, Misra, & Hebert, 2017), which natural blurring is similar effect to Poisson blending. We synthesized two sets of anomaly images. One of the two sets includes 2801 images. We localized an anomaly object on the images manually. The other set includes 2836 images. Positions of anomaly objects are selected randomly from undamaged regions.

**Metrics**: We use two metrics for evaluating each model in our ensemble model. Mean average precision (AP) evaluates supervised RDD. Pixel level Area Under Curve(AUC) evaluates anomaly detection model.

### 4.2. Results

#### 4.2.1. Anomaly Detection

Before evaluating our ensemble model, we show evaluations of each single model in our ensemble model. The first evaluation confirms performances of the recent anomaly detection

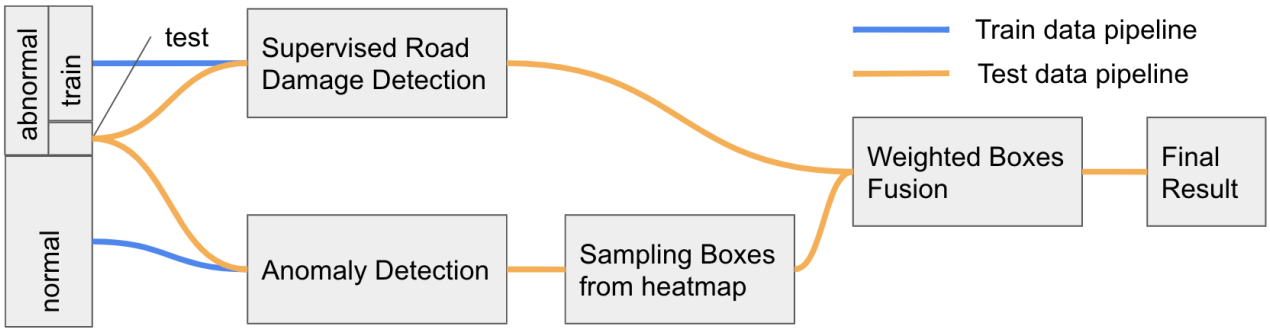


Figure 1. Pipeline of our ensemble model. Our model is consisted of anomaly detection and supervised road damage detection.

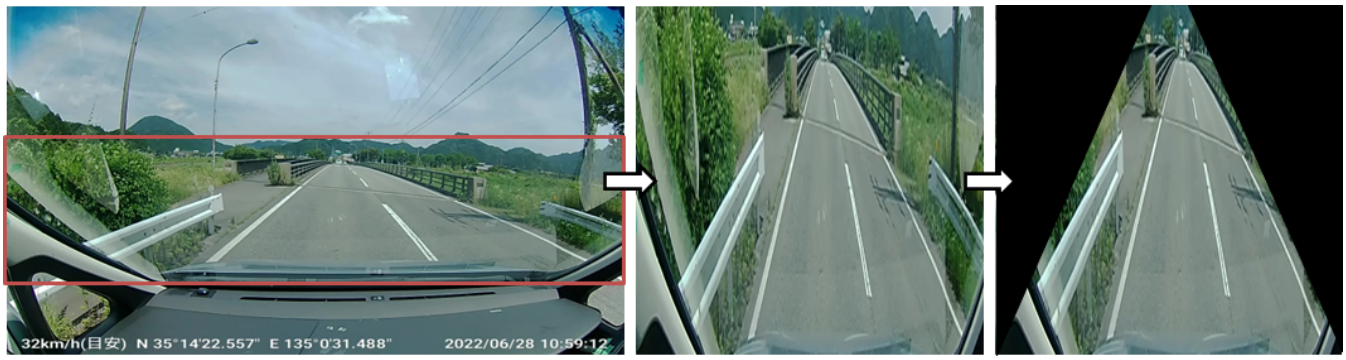


Figure 2. Pictorial cropping pipeline for eliminating roadside objects.

Figure 3. Road damages and their segmentation masks. From the left, the first and the second images show a linear crack and the segmentation mask, respectively. The third and the forth show a pothole and the segmentation mask.

models, when trapezoidal cropping is applied as preprocessing. The result is shown in Table 2. We noticed that trapezoidal cropping improved the results of PaDiM and CSFlow. However, the cropping deteriorated performance of CFA (Lee et al., 2022). We considered that cropped regions may harm a model of CFA, because, unlike other methods, CFA focuses the adapting patch descriptors to the target domain. The high simplicity of the cropped regions in comparison to the normal image features acquired through pre-training may lead to a higher relative anomaly in the normal regions. This may be caused by the contrastive learning in CFA, which lacks an absolute reference point. In our experiment, PaDiM achieves the highest pixel level AUC.

Figure 4 depicts the qualitative inference result of PaDiM. The first column shows the original images, the second is the ground truth, the third is the estimated anomaly heatmap, the fourth is the thresholded segmentation result, and the fifth is the detected anomaly regions on the images. We observed numerous false positives on areas of the image that are not abnormal. In the images in the second row, multiple false positives are present on a normal road surface, even though several damages can be detected correctly. The third row shows false positives on both sides of the road. PaDiM mistakenly detects shades, a vehicle, and glasses as anomaly objects.

From the above evaluation, we considered that a single anomaly detection model cannot reduce false positives. An eminent effect of anomaly detection is finding image patterns that are rarely observed on a road. When aiming to detect all road damages by a single anomaly detection model, we must collect all normal road patterns on a road. In addition to existing normal patterns, we have to handle wide varieties of shades by roadside objects, such as trees, road signs, and utility poles. Since shades vary by light condition of the sun, diverse light conditions should be concerned when collecting normal images. Therefore, we concluded that only an anomaly detection model cannot overcome to suppress these false positives.

#### 4.2.2. Supervised Road Damage Detection

As well as evaluations of anomaly detection, we evaluate a single supervised RDD model that includes in our ensemble model. Table 1 presents performance comparison between

several CNNs used as supervised RDD. The CNNs are selected to implement on an embedding machine(i.e., working on a poor computational resource). (a) in Table1 shows the results evaluated on one of synthetic datasets, which positions of anomaly objects are directed manually. We considered that every model can detect damages successfully, as the most of APs in (a) surpass 90. This is because both training and test datasets include the same feature patterns of damage.

(b) in Table 1 shows the results evaluated on another synthetic dataset, where the positions of anomaly objects are randomly selected from undamaged regions. We cannot observe a large gap between (a) and (b) in Table 1, because most of the APs exceed 85. These results indicate that supervised RDD do not strongly depend on light conditions, roadside objects, and other vehicles. In other words, supervised RDD can learn feature patterns of damages on roads.

We elaborately evaluate the CNNs on another dataset, which damages are naturally occurred on actual roads. The result is shown in (c) of Table1. We found that performances of all models are drastically down compared with the above results on synthetic datasets. In (c) of Table1, AP of "closed" row is higher than other rows. A model of "closed" row indicates that trainig and test datasets are the same, completely. The result means that supervised RDD could not train wide varieties of road features in actual scenes. Even with the current state-of-the-art of light CNN, AP50val in Table1 is under 30.

#### 4.2.3. Ensemble model & Discussion for future work

We evaluate our ensemble model shown in Figure 1. The model consists of supervised RDD and anomaly detection. Table 3 shows the APs of our ensemble model. The table reveals that the AP of our ensemble model is lower than that of a single supervised RDD. This result is attributed to the occurrence of false positives in the anomaly detection component of our ensemble model.

We hypothesized that the negative result may be caused by limitations of fusion. Weighted boxes fusion in our ensemble model have several drawbacks, such as difficulty of hyperparameter tuning, even treatments to all fused models, and disregard for small objects. In addition to fusion, we considered that preprocessing is another cause of the result. We can't efficiently suppress distractions (e.g., light condition, objects except for roads). Semantic road segmentation is a candidate to resolve the issue, even if the approach mistakenly segments road region.

## 5. Conclusion

Toward road damage detection that finds also anomaly phenomena on roads, we proposed an ensemble model of supervised road damage detection and anomaly detection. First, we evaluated a single model in our ensemble model on our driving recorder dataset. For avoiding performance degradation by roadside objects, we applied trapezoidal cropping as pre-processing. Evaluations showed that supervised RDD works well on synthetic datasets and degrades on an actual dataset. As anomaly detection, we could observe false positives caused by remaining roadside objects.

Second, we proposed an ensemble model to aim better performance than a single model. Unfortunately, we couldn't achieve better result. We considered that further research is required to fuse multiple models for better performance.

## Acknowledgment

We express our gratitude to the members of the Mitsubishi Electric Mobility Service Development Division's Business Strategy Project for their support in data acquisition through the use of the drive recorder. We would also like to extend our thanks to committee member Hidetoshi Mishima.

## References

Akcay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2019). Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer vision–accv 2018: 14th asian conference on computer vision, perth, australia, december 2–6, 2018, revised selected papers, part iii 14* (pp. 622–637).

Bukhsh, Z. A., Jansen, N., & Saeed, A. (2021). Damage detection using in-domain and cross-domain transfer learning. *Neural Computing and Applications*, *33*(24), 16921–16936.

Defard, T., Setkov, A., Loesch, A., & Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern recognition. icpr international workshops and challenges: Virtual event, january 10–15, 2021, proceedings, part iv* (pp. 475–489).

Dwibedi, D., Misra, I., & Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the ieee international conference on computer vision* (pp. 1301–1310).

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Hegde, V., Trivedi, D., Alfarrarjeh, A., Deepak, A., Kim, S. H., & Shahabi, C. (2020). Yet another deep learning approach for road damage detection using ensemble learning. In *2020 ieee international conference on big data (big data)* (pp. 5553–5558).

Jo, Y., & Ryu, S. (2015). Pothole detection system using a black-box camera. *Sensors*, *15*(11), 29316–29331.

Lee, S., Lee, S., & Song, B. C. (2022). Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, *10*,

Table 1. Mean Average Precision of a single supervised RDD. Subscripts of AP are defined by MSCOCO evaluation tools. A shape of all damage images are square. S: an edge of a damage image is shorter than 32 pixels. M: the same edge is 32 to 96 pixels. L: the same edge is longer than 96 pixels. The term "closed" in (c) indicates that the performance of the corresponding row is evaluated under the same dataset for both training and testing.

(a) Our Drive Recorder Dataset (synthesized damage, fixed position)

| **Model** | **Params** | $\mathbf{AP}^{\mathbf{val}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{50}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{75}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{S}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{M}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{L}}$ |
|---|---|---|---|---|---|---|---|
| YOLOX-s | 9.0 M | 90.9 | 97.9 | 96.3 | 77.9 | 89.5 | 94.7 |
| YOLOX-l | 54.2 M | 92.7 | 98.1 | 96.8 | 80.6 | 92.0 | 95.5 |
| YOLOv7-tiny | 6.2 M | 88.0 | 96.8 | 94.9 | 75.2 | 85.7 | 92.5 |
| YOLOv7 | 36.9 M | 91.7 | 97.3 | 95.5 | 77.6 | 90.9 | 94.6 |
| YOLOv8-s | 11.2 M | 95.4 | 97.3 | 96.7 | 84.9 | 95.2 | 97.0 |
| YOLOv8-l | 43.7 M | 96.2 | 97.9 | 96.9 | 87.4 | 96.0 | 96.6 |

(b) Our Drive Recorder Dataset (synthesized damage, random position)

| **Model** | **Params** | $\mathbf{AP}^{\mathbf{val}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{50}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{75}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{S}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{M}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{L}}$ |
|---|---|---|---|---|---|---|---|
| YOLOX-s | 9.0 M | 88.3 | 99.0 | 97.4 | 77.4 | 84.1 | 94.9 |
| YOLOX-l | 54.2 M | 85.6 | 98.9 | 97.0 | 73.0 | 81.6 | 92.3 |
| YOLOv7-tiny | 6.2 M | 87.3 | 97.3 | 95.4 | 77.4 | 82.2 | 93.7 |
| YOLOv7 | 36.9 M | 88.7 | 97.7 | 96.2 | 75.3 | 84.4 | 95.5 |
| YOLOv8-s | 11.2 M | 96.6 | 99.0 | 98.5 | 88.4 | 95.4 | 98.6 |
| YOLOv8-l | 43.7 M | 97.1 | 99.5 | 98.5 | 89.2 | 95.8 | 98.8 |

(c) Our Drive Recorder Dataset (actual damage)

| **Model** | **Params** | $\mathbf{AP}^{\mathbf{val}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{50}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{75}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{S}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{M}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{L}}$ |
|---|---|---|---|---|---|---|---|
| YOLOX-s | 9.0 M | 9.3 | 27.0 | 4.3 | 3.2 | 6.6 | 13.5 |
| YOLOX-l | 54.2 M | 10.4 | 27.2 | 7.2 | 3.5 | 6.4 | 17.0 |
| YOLOv7-tiny | 6.2 M | 14.0 | 35.8 | 8.6 | 3.4 | 11.7 | 21.4 |
| YOLOv7 | 36.9 M | 15.9 | 38.4 | 12.2 | 4.0 | 11.5 | 24.7 |
| YOLOv8-s (closed) | 11.2 M | 64.5 | 83.1 | 71.4 | 11.8 | 71.3 | 87.7 |
| YOLOv8-s | 11.2 M | 15.9 | 33.1 | 14.2 | 5.9 | 8.6 | 23.8 |
| YOLOv8-l | 43.7 M | 15.9 | 35.4 | 12.1 | 4.4 | 10.8 | 25.3 |

Table 2. Pixel-level AUC of anomaly detection. "w/o" indicates that trapezoidal cropping is not applied at pre-processing.

| **Model** | Backbone | $\mathbf{w/o}$ | $\mathbf{w/}$ |
|---|---|---|---|
| CFA | Resnet-18 | 0.71 | 0.47 |
| CFA | Wide Resnet-50 | 0.67 | 0.54 |
| PaDiM | Resnet-18 | 0.70 | 0.79 |
| PaDiM | Wide Resnet-50 | 0.67 | 0.80 |
| CSFlow | | 0.57 | 0.66 |

78446–78454.

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., & Omata, H. (2018). Road damage detection and classification using deep neural networks with smartphone images. *Computer-Aided Civil and Infrastructure Engineering*, *33*(12), 1127–1141.

Nienaber, S., Booysen, M. J., & Kroon, R. (2015). Detecting potholes using simple image processing techniques and real-world footage.

Rudolph, M., Wehrbein, T., Rosenhahn, B., & Wandt, B. (2022). Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1088–1097).

Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, *107*, 104117.

Tao, X., Zhang, D., Wang, Z., Liu, X., Zhang, H., & Xu, D. (2018). Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(4), 1486–1498.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2021). You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*.

Xu, J., Xiong, Z., & Bhattacharyya, S. P. (2022). Pidnet: A real-time semantic segmentation network inspired from pid controller. *arXiv preprint arXiv:2206.02066*.

Zhao, H., Zhang, H., & Zhao, Y. (2023). Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 233–238).

Table 3. Comparison between supervised RDD and ensemble models. Both models are evaluated on actual damage dataset

| **Model** | $\mathbf{AP}^{\mathbf{val}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{50}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{75}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{S}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{M}}$ | $\mathbf{AP}^{\mathbf{val}}_{\mathbf{L}}$ |
|---|---|---|---|---|---|---|
| YOLOv8-s | 15.9 | 33.1 | 14.2 | 5.9 | 8.6 | 23.8 |
| YOLOv8-s + PaDiM | 2.7 | 5.0 | 2.8 | 1.4 | 1.5 | 6.2 |
| YOLOv8-l | 15.9 | 35.4 | 12.1 | 4.4 | 10.8 | 25.3 |
| YOLOv8-l + PaDiM | 2.1 | 4.1 | 2.0 | 0.0 | 0.0 | 9.6 |

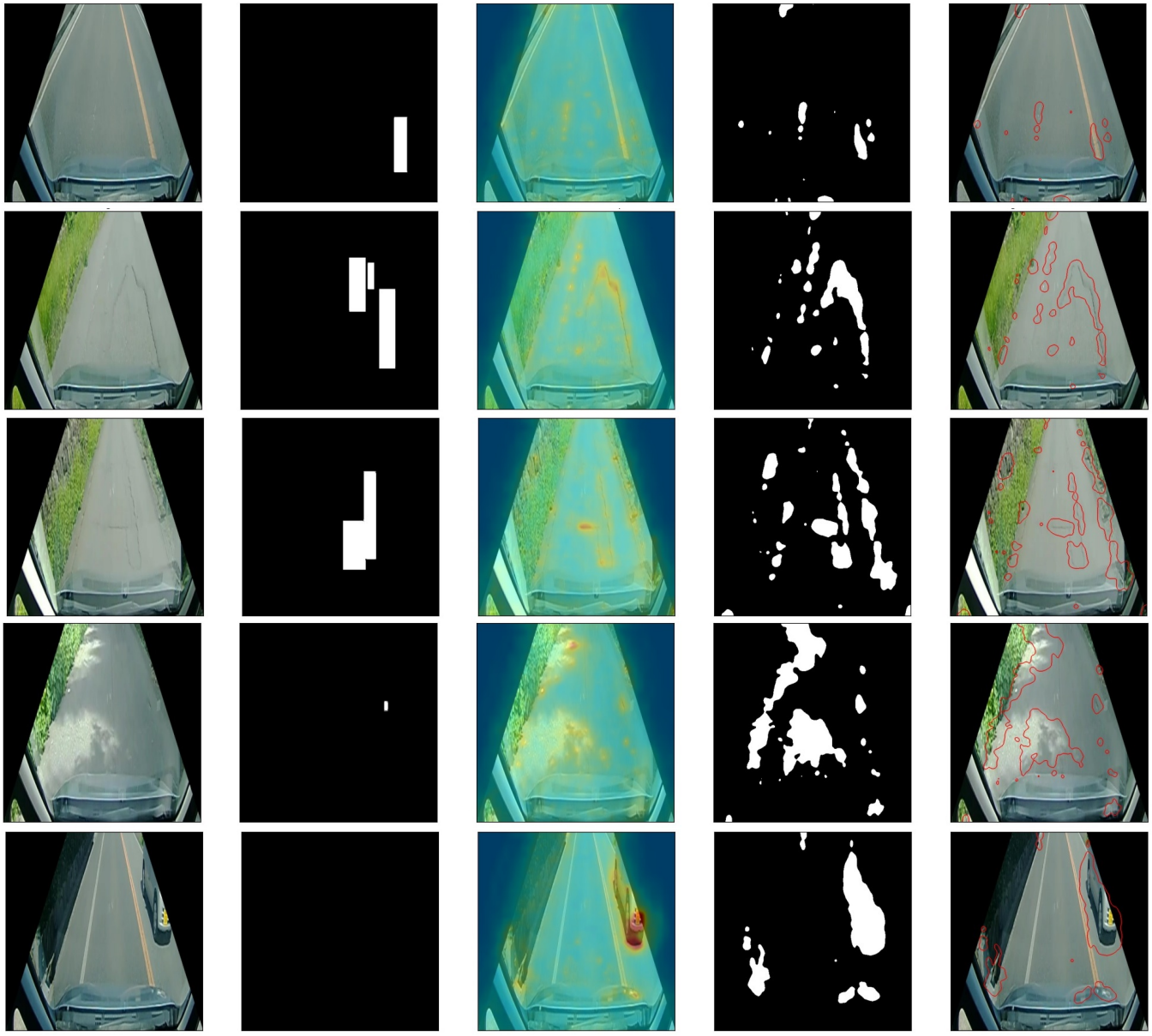

Figure 4. Qualitative inference results of PaDiM