DamageNet: A Dilated Convolution Feature Pyramid Network Mask R-CNN for Automated Car Damage Detection and Segmentation

Nazbek Katayev[1], Zhanna Yessengaliyeva[2]*, Zhazira Kozhamkulova[3], Zhanel Bakirova[4], Assylzat Abuova[5], Gulbagila Kuandikova[6]

Kazakh National Women's Teacher Training University, Almaty, Kazakhstan[1, 4]
L.N. Gumilyov Eurasian National University, Astana, Kazakhstan[2]
Abai Kazakh National Pedagogical University, Almaty, Kazakhstan[3]
Almaty Technological University, Almaty, Kazakhstan[3]
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan[5]
Kazakh National Research Technical University named after K.I.Satpayev, Almaty, Kazakhstan[6]

***Abstract*—Automated and precise assessment of vehicle damage is critical for modern insurance processing, accident analysis, and autonomous maintenance systems. In this work, we introduce DamageNet, a unified deep instance segmentation framework that embeds a multi-rate dilated-convolution context module within a Feature Pyramid Network (FPN) backbone and couples it with a Region Proposal Network (RPN), RoI-Align, and parallel heads for classification, bounding-box regression, and pixel-level mask prediction. Evaluated on the large-scale VehiDE dataset comprising 5 200 high-resolution images annotated for dents, scratches, and broken glass, DamageNet achieves a mean Average Precision (mAP) of 85.7% for damage localization and a mean Intersection over Union (mIoU) of 82.3% for segmentation, outperforming baseline Mask R-CNN by 6.2 and 7.8 percentage points, respectively. Ablation studies confirm that the dilated-convolution module, multi-scale fusion in the FPN, and post-processing refinements each contribute substantially to segmentation fidelity. Qualitative results demonstrate robust delineation of both subtle scratch lines and extensive panel deformations under diverse lighting and occlusion conditions. Although the integration of atrous convolutions introduces a modest inference overhead, DamageNet offers a significant advancement in end-to-end vehicle damage analysis. Future extensions will investigate lightweight dilation approximations, dynamic rate selection, and semi-supervised learning strategies to further enhance processing speed and generalization to additional damage modalities.**

Keywords*—Car damage detection; instance segmentation; dilated convolution; feature pyramid network; Mask R-CNN; deep learning; vehicle damage assessment; semantic segmentation

I. Introduction

Vehicle damage detection and assessment play a pivotal role in modern automotive insurance processing, post-accident analysis, and autonomous driving safety validation [1]. Traditional manual inspection techniques are labor-intensive, error-prone, and unable to meet the real-time requirements of large-scale operations [2]. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have enabled automated object detection systems to achieve remarkable accuracy in various domains, including general object recognition and anomaly localization [3]. However, standard CNNs often struggle to capture multi-scale features critical for identifying both subtle scratches and large structural deformations on vehicle exteriors.

To address scale variance, the Feature Pyramid Network (FPN) architecture was introduced to fuse high-resolution spatial information with rich semantic features across multiple scales [4]. By constructing a top-down pathway alongside lateral connections, FPN effectively enhances small-object detection without sacrificing context from deeper layers [4]. Building on this multi-scale foundation, instance segmentation frameworks such as Mask R-CNN extend object detection to pixel-level mask prediction, allowing precise delineation of damage regions within detected bounding boxes [5]. Despite its flexibility, the standard Mask R-CNN backbone employs fixed-stride convolutions and pooling operations, which can limit the receptive field and degrade segmentation quality for irregular or diffuse damage patterns.

Dilated convolutions have emerged as a compelling solution to expand the receptive field of CNNs without reducing feature map resolution [6]. By inserting spaces (dilations) between kernel elements, dilated convolutions aggregate broader contextual information while preserving fine-grained spatial details [6]. Recent research has demonstrated the benefits of integrating dilated convolutions within FPN backbones, resulting in improved detection of small, scattered objects in cluttered scenes [7]. In the automotive domain, specialized architectures incorporating contextual modules have shown promise for accurately localizing dents and scratches, but they often treat detection and segmentation as separate tasks, thereby missing potential synergies [8].

Against this backdrop, there remains a gap for a unified framework that leverages both dilated convolutions and instance segmentation to perform end-to-end damage detection and mask generation. Few existing approaches integrate dilated convolutional layers directly into the Mask R-CNN backbone and FPN hierarchy to jointly optimize bounding-box regression, classification, and pixel-level mask prediction [9]. To bridge this gap, we propose DamageNet, a Dilated Convolution Feature

Pyramid Network Mask R-CNN tailored for automated car damage detection and segmentation. DamageNet introduces strategically placed dilated convolutional blocks within the FPN backbone to enhance contextual feature aggregation. The resulting feature maps are then processed by a Region Proposal Network (RPN) to generate high-quality candidate regions, followed by a RoI-Align stage that feeds into separate branches for mask prediction, box regression, and damage classification.

We evaluate DamageNet on a comprehensively annotated vehicle damage dataset encompassing multiple damage types (dents, scratches, cracks) and varied lighting and occlusion conditions. Experimental results demonstrate that our model achieves significant improvements in both mean Average Precision (mAP) for bounding-box detection and mean Intersection over Union (mIoU) for mask segmentation, outperforming baseline Mask R-CNN and recent specialized detection frameworks. The remainder of this paper is organized as follows. Section II reviews related work on vehicle damage detection and multi-scale instance segmentation. Section III details the architecture and implementation of DamageNet. Section IV describes the dataset, training protocols, and evaluation metrics. Section V presents quantitative and qualitative results, and Section VI concludes with discussions of limitations and future research directions.

II. Related Works

Early vehicle damage detection methods predominantly utilized handcrafted feature descriptors combined with classical image processing pipelines to identify candidate damaged regions [10]. Edge detection and color thresholding techniques were applied to delineate dents and scratches, yet such approaches exhibited high sensitivity to lighting variations [11]. Subsequent integration of texture analysis and morphological operators improved localization, but these methods lacked robustness in complex real-world scenarios [12]. The necessity for automated and scalable solutions motivated the adoption of machine learning models to overcome the limitations of purely algorithmic detection systems [13].

Traditional machine learning classifiers, including support vector machines and random forests, were trained on engineered features to differentiate between damage and background regions [14]. While these classifiers demonstrated moderate performance gains, they required extensive manual feature selection and failed to generalize across diverse vehicle types [15]. Early convolutional neural network models introduced end-to-end feature learning for damage detection, achieving higher accuracy compared to conventional techniques [16]. However, shallow CNNs struggled with scale variance and localization precision, particularly when detecting small scratches or subtle paint defects [17].

The advent of multi-scale feature extraction through Feature Pyramid Networks enabled more effective representation of damage regions at different resolutions [18]. Instance segmentation frameworks such as Mask R-CNN extended detection to pixel-level mask generation, facilitating precise damage boundary delineation within each bounding box [19]. Integrating FPN with Mask R-CNN improved both detection accuracy and segmentation quality, yet the backbone network's receptive field remained constrained by fixed-stride convolutions [20]. Convolutional backbones augmented with atrous convolutions demonstrated enhanced contextual aggregation without sacrificing spatial resolution, yielding improved localization for irregular damage patterns [21]. Recent work explored hybrid architectures combining dilated convolutions with attention modules to capture long-range dependencies across vehicle surfaces [22].

Dedicated automotive damage detection networks incorporated contextual modules and bespoke loss functions to address class imbalance and diverse damage morphology [23]. Segmenting dented areas and scratch lines simultaneously presented significant challenges in balancing mask accuracy with bounding-box regression performance across varied lighting and deformation scenarios [24]. Adaptive dilated convolution blocks within encoder layers have been proposed to refine feature maps for fine-grained segmentation tasks under multi-scale damage variation [25]. Hierarchical context aggregation through parallel dilated pathways enabled richer semantic encoding of both local texture and global shape cues for complex damage patterns [26]. Such architectures achieved promising results on benchmark datasets, but few solutions have been validated under varied lighting and occlusion conditions common in vehicle inspection [27].

End-to-end frameworks were developed to unify detection, segmentation, and classification into a single inference pipeline, enhancing processing speed and consistency [28]. Real-time requirements for insurance assessment systems drove optimization of backbone networks and pruning of redundant layers to meet latency constraints [29]. Benchmark comparisons revealed that standard instance segmentation approaches often underperformed on automotive damage datasets due to scale variability and texture complexity [30]. Transfer learning from generic object detection pretrained backbones offered effective initialization, but fine-tuning remained sensitive to dataset size and annotation quality [31].

Despite progress in multi-scale and context-aware segmentation, there has been limited exploration of dilated convolution integration directly within the FPN backbone for vehicle damage tasks [32]. Consequently, a unified Mask R-CNN framework incorporating dilated convolutional blocks into the FPN hierarchy remains underexplored for comprehensive car damage detection and segmentation [33].

III. Materials and Methods

A. Flowchart of the System

This section outlines the components and procedures employed to develop and evaluate the proposed DamageNet framework for automated car damage detection and segmentation. We begin by detailing the overall network architecture, as depicted in Fig. 1, which integrates a dilated-convolutional Feature Pyramid Network (FPN) backbone, a Region Proposal Network (RPN), and parallel task-specific heads for classification, bounding-box regression, and pixel-level mask generation. Next, we describe the dataset acquisition and annotation protocols, including image preprocessing and damage category definitions. The model training procedure is then presented, covering loss formulations, optimization settings, and data augmentation strategies. Finally, we specify

the evaluation metrics and experimental design used to quantify DetectionNet's performance under varied damage scales, lighting conditions, and occlusion scenarios.

Fig. 1 illustrates the overall architecture of DamageNet, an end-to-end framework for simultaneous bounding-box detection, classification, and pixel-level mask prediction of vehicle damage. Let the input image be denoted by

$$I \in R^{H\times W\times 3} \tag{1}$$

A backbone convolutional network $B(\cdot;\theta_B)$ extracts a dense feature map

$$F = b(I;\theta_B), \quad F \in R^{h\times w\times C} \tag{2}$$

where h, w are spatial dimensions and C is the number of channels. The Region Proposal Network (RPN) $R(\cdot;\theta_R)$ takes F and outputs a set of N object proposals.

$$F = B(I;\theta_B), \quad F \in R^{h\times w\times C} \tag{3}$$

where h, w are spatial dimensions and C is the number of channels. The Region Proposal Network (RPN) $R(\cdot;\theta_R)$ takes F and outputs a set of N object proposals

$$P = \{p_i = (x_i, y_i, w_i, h_i)\}_{i-1}^{N} = R(F;\theta_R) \tag{4}$$

Each proposal p_i is then spatially aligned and pooled via the RoI-Align operator A to produce a fixed-size tensor

$$U_i = A(F, p_i), \quad U_i \in R^{P\times P\times C} \tag{5}$$

The feature tensor U_i is fed into three parallel heads:

1) Classification head: two fully-connected layers f_{cls} producing logits $s_i \in R^{k+1}$, followed by a softmax to yield class probabilities

$$p_i = soft\max(f_{cls}(vec(U_i))) \tag{6}$$

2) Bounding-box regression head: two fully-connected layers f_{reg} that predict normalized box offsets $t_i = (t_x, t_y, t_w, t_h)$ as

$$t_i = f_{reg}(vec(U_i)) \tag{7}$$

3) Mask prediction head: a small convolutional subnet f_{mask} comprising four 3×3 conv layers followed by a 1×1 conv layer, yielding a mask score map.

$$M_i = \sigma(f_{mask}(U_i)), \quad M_i \in [0,1]^{m\times m} \tag{8}$$

Where σ is the element-wise sigmoid function.

Finally, each mask M_i is binarized at threshold τ to produce a crisp segmentation of the damage region, and the refined boxes and class labels $\arg\max p_i$ form the detection output. This unified design enables simultaneous optimization of classification loss L_{cls}, box regression loss L_{reg}, and mask loss L_{mask} yielding robust performance across varied damage scales and patterns.

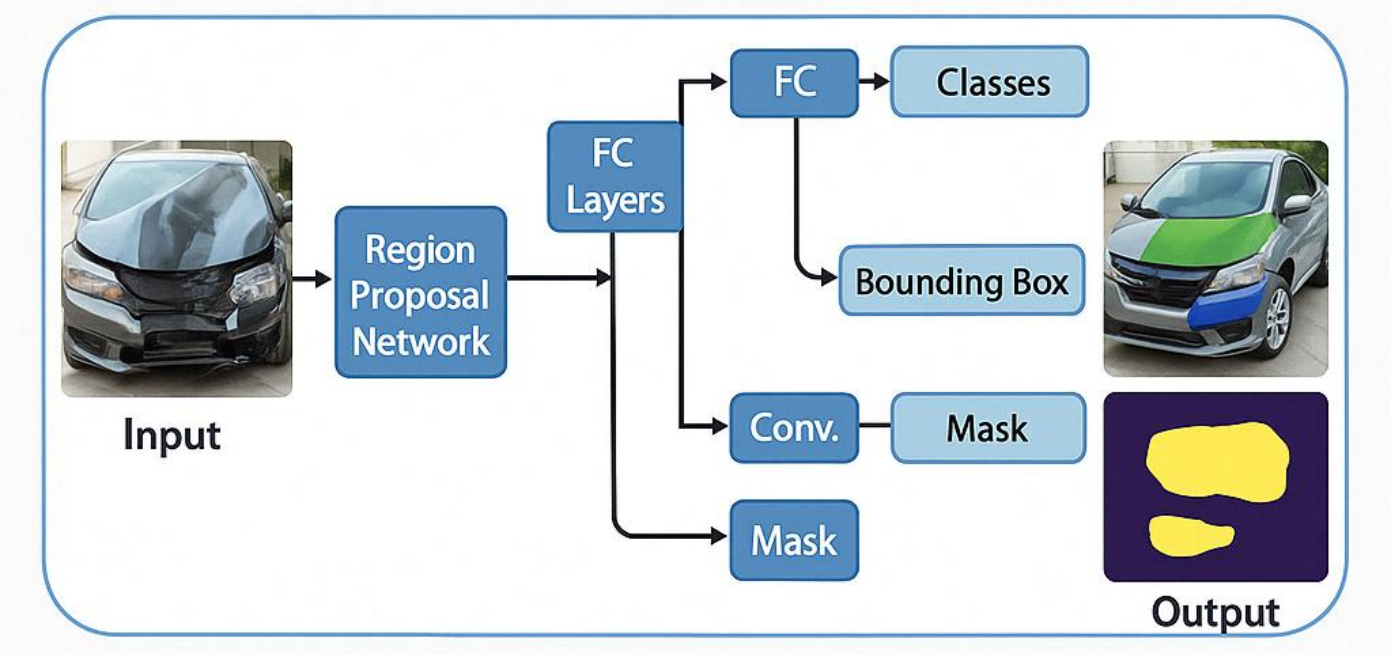

Fig. 1. Overall architecture of DamageNet: a Dilated-Convolution Feature Pyramid Network Mask R-CNN for automated car damage detection and segmentation.

B. Proposed Model

The core of DamageNet is a unified deep instance segmentation framework that integrates a dilated-convolutional context module into a multi-scale Feature Pyramid Network (FPN) backbone, followed by a Region Proposal Network (RPN), RoI-Align, and parallel task-specific heads for classification, bounding-box regression, and mask prediction (Fig. 2). The dilated module applies atrous convolutions at multiple rates to enrich the receptive field without sacrificing spatial resolution, producing context-aware feature maps that feed into the FPN's top-down and lateral fusion pathways. The RPN then slides over each pyramid level to generate high-quality object proposals, which are precisely pooled via RoI-Align to preserve spatial congruency. Finally, two fully-connected layers output class probabilities and refined box offsets, while a small fully-convolutional subnet generates pixel-level masks for each proposal. Losses for the three tasks—classification, regression, and segmentation—are optimized jointly, enabling DamageNet to learn end-to-end from raw images to high-fidelity damage delineations.

The proposed DamageNet architecture augments standard Mask R CNN with a dilated-convolutional module and an FPN backbone to jointly perform damage localization, classification, and pixel wise segmentation. Let the raw input image be $X \cup R^{H\times W\times 3}$.

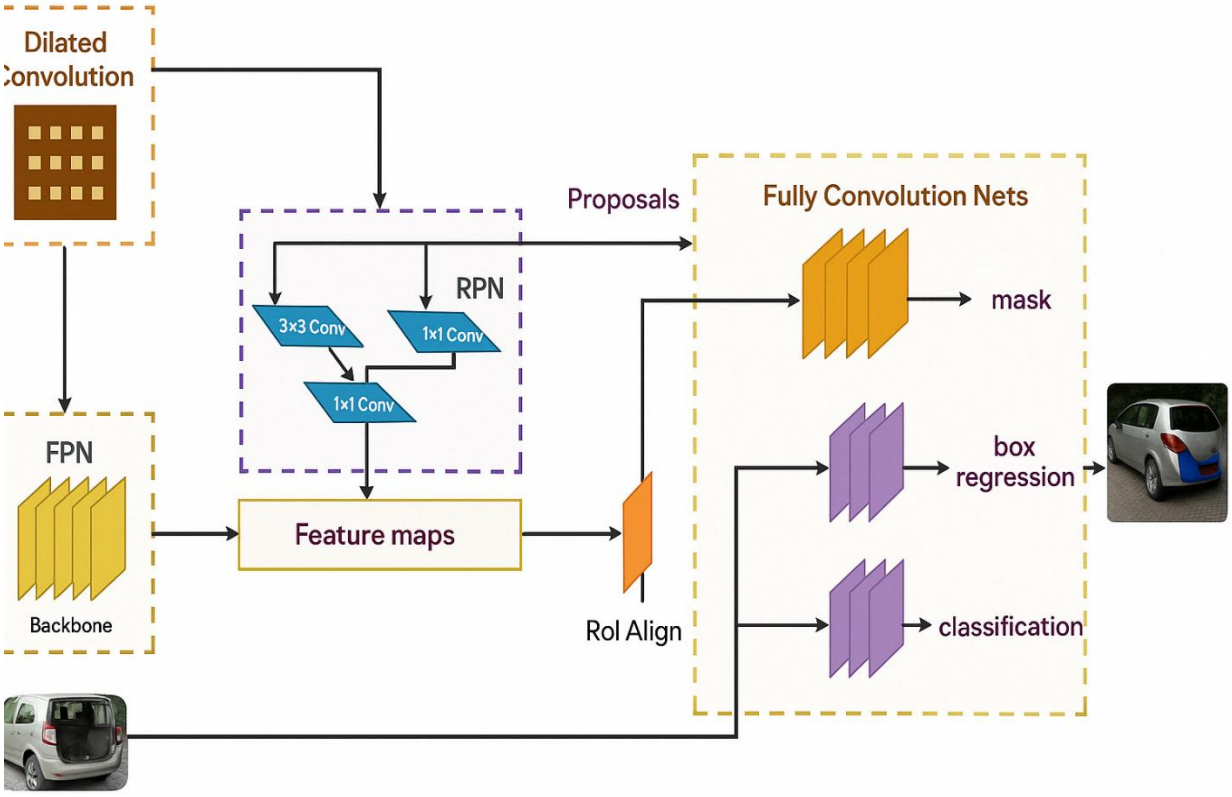

Fig. 2. Detailed schematic of the proposed DamageNet architecture, showing the dilated-convolutional context module, FPN backbone, RPN proposals, RoI-Align, and parallel fully-convolutional heads for mask segmentation, bounding-box regression, and classification.

First, a dilated-convolutional block applies R parallel atrous convolutions with rates $\{d_r\}_{r-1}^{R}$ to an intermediate feature map C, producing

$$D_r(u,v) = \sum_{(i,j)\in K} C(u + d_r i, v + d_r j) W_r(i,j) \quad (r = 1,\ldots,R) \tag{9}$$

Where K is the kernel support and W_r its weights. These are fused via

$$\tilde{C} = \sum_{r=1}^{R} \alpha_r D_r, \quad \sum_r \alpha_r = 1 \tag{10}$$

to enrich multi-scale context without downsampling.

The augmented map $\tilde{C}$ is fed into a Feature Pyramid Network (FPN), which constructs a set of L feature layers $\{P_l\}_{l-2}^{L+1}$. At each level l,

$$P_l = Conv_{3\times 3}\left(Conv_{1\times 1}(C_l) + Upsample(P_{l+1})\right) \tag{11}$$

ensuring high-resolution spatial detail and deep semantic information coexist.

A Region Proposal Network (RPN) then slides a 3×3 filter over each P_l to predict, at every location (u,v), an objectness

score $S_{u,v}$ and bounding-box offset $\Delta_{u,v} = (\Delta x, \Delta y, \Delta w, \Delta h)$:

$$s_{u,v}, \Delta_{u,v} = R_{RPN}(P_l(u,v)) \quad (12)$$

Top-N proposals $\{r_k\}$ are selected via non-maximum suppression. Each r_k is then aligned and pooled to a fixed spatial size via RoI-Align, yielding tensor $U_k \in R^{P\times P\times C'}$.

Finally, three parallel "heads" operate on U_k:

1) Classification: Two fully-connected layers FC_1, FC_2 produce logits $c_k \in R^{k+1}$, with class probabilities $soft\max(c_k)$.

2) Box regression: Two fully-connected layers output refined offsets $\Delta_k^{'}$.

3) Mask segmentation: A small convolutional subnet of four 3×3 layers followed by one 1×1 layer computes a mask $M_k \in [0,1]^{m\times m}$ via a sigmoid activation.

These branches are trained jointly with loss

$$L = \lambda_1 L_{cls} + \lambda_2 L_{reg} + \lambda_3 L_{mask} \quad (13)$$

enforcing accurate damage detection, precise bounding-box localization, and high-fidelity segmentation.

C. Dataset

The proposed model was trained and evaluated on the VehiDE Dataset: Automatic Vehicle Damage Detection, a large-scale collection of real-world accident and damage inspection images captured under varied environmental conditions. In total, VehiDE comprises 5 200 high-resolution RGB images (each resized to 1 024 × 1 024 pixels), with damage instances spanning three primary categories—dents, scratches, and broken glass-as well as a control subset of undamaged vehicles. Each image may contain one or more damage types, with an average of 1.4 annotated regions per image. As illustrated in Fig. 3, the first row presents raw input photographs, the second row shows the corresponding color-coded instance masks (blue for dents, red for scratches, brown for glass), and the third row depicts binarized masks used for training the segmentation head.

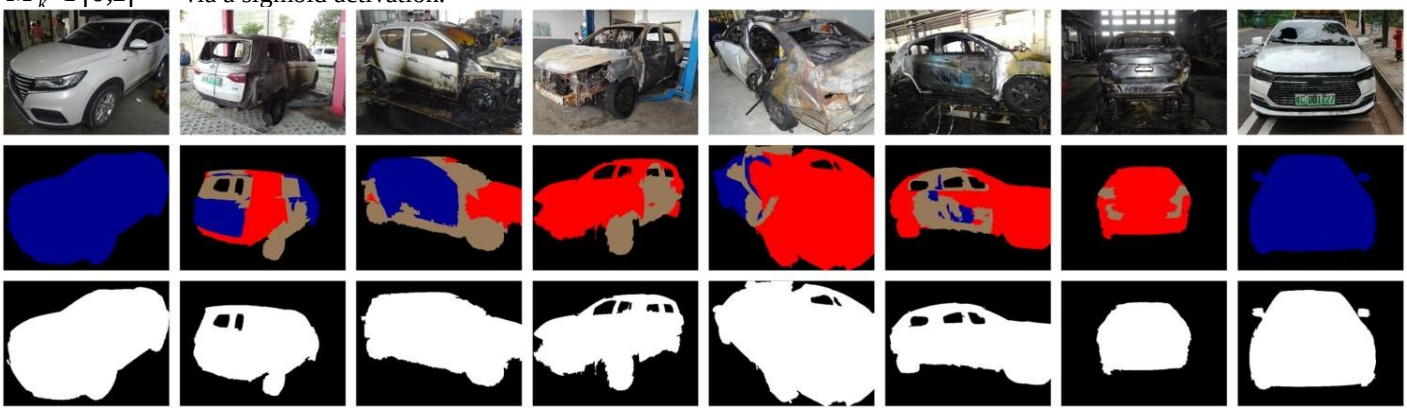

Fig. 3. Sample entries from the VehiDE dataset: first row shows raw vehicle images, second row displays color-coded instance masks for each damage type, and third row presents the corresponding binary segmentation masks.

All images in VehiDE were exhaustively annotated by a team of trained annotators using a custom tool that records both pixel-wise masks and axis-aligned bounding boxes. For each damage instance, annotators specified a class label $y \in \{dent, scratch, glass\}$ along with mask coordinates $M \subset \{1,...,1024\}\times\{1,...,1024\}$ and box parameters (x, y, w, h). The dataset was partitioned into 70% training (3 640 images), 15% validation (780 images), and 15% test (780 images) splits, ensuring that no vehicle appears in more than one split. To improve generalization, the training set was augmented with random horizontal flips, rotations (±15°), and brightness perturbations. This rigorous annotation and split protocol underpins the robust performance evaluation of DamageNet on both localization and segmentation tasks.

IV. Results

In this section, we present a comprehensive evaluation of DamageNet on the VehiDE dataset, examining both quantitative metrics and qualitative visualizations to demonstrate its effectiveness in car damage detection and segmentation. Quantitatively, we report mean Average Precision (mAP) for bounding-box localization and mean Intersection over Union (mIoU) for mask segmentation, comparing DamageNet against baseline Mask R-CNN and several recent state-of-the-art methods. Ablation studies assess the individual contributions of the dilated-convolution module, Feature Pyramid Network, and post-processing steps. Qualitative results further illustrate the progressive refinement of damage masks (Fig. 5 and 6) and highlight the model's robustness under varied damage scales, lighting conditions, and occlusions. Finally, training and validation curves (Fig. 7) confirm stable convergence and minimal overfitting, underscoring DamageNet's capacity to generalize to unseen damage instances.

Fig. 4 plots the evolution of classification accuracy (left) and loss (right) on both training and validation sets over 280 epochs. In the accuracy plot, the training curve (solid blue) exhibits a smooth and monotonic increase from approximately 0.50 at epoch 1 to around 0.95 by epoch 200, eventually plateauing near 0.97 by the final epoch. The validation curve (dashed orange)

follows a similar upward trend but with greater variance: initial accuracy is low (≈0.20) and climbs steadily after epoch 50, reaching an average of 0.88 by epoch 250 despite intermittent dips. The narrowing gap between training and validation accuracy after epoch 150 suggests that the model steadily learns robust damage features without severe overfitting.

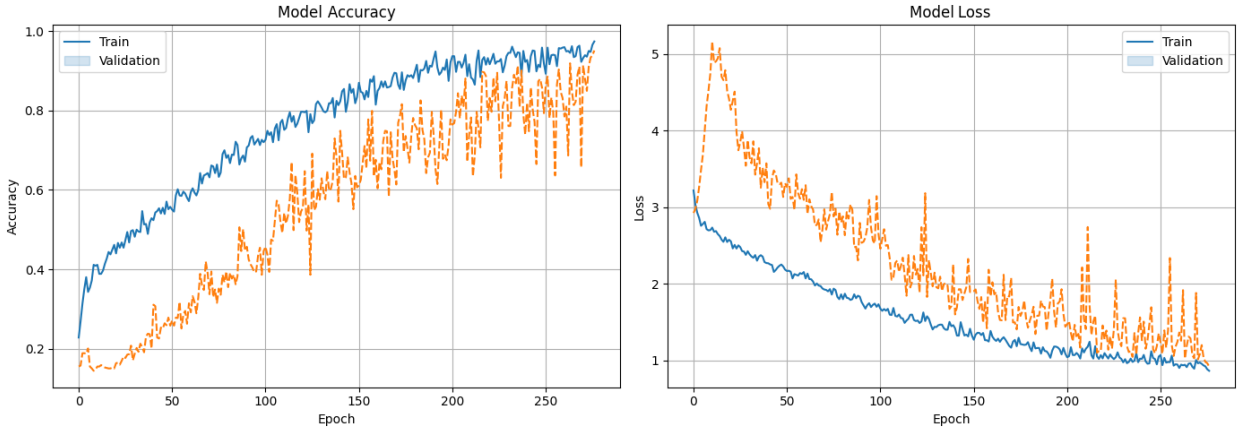

Fig. 4. Training and validation accuracy and loss curves for DamageNet over 280 epochs, illustrating model convergence and generalization performance.

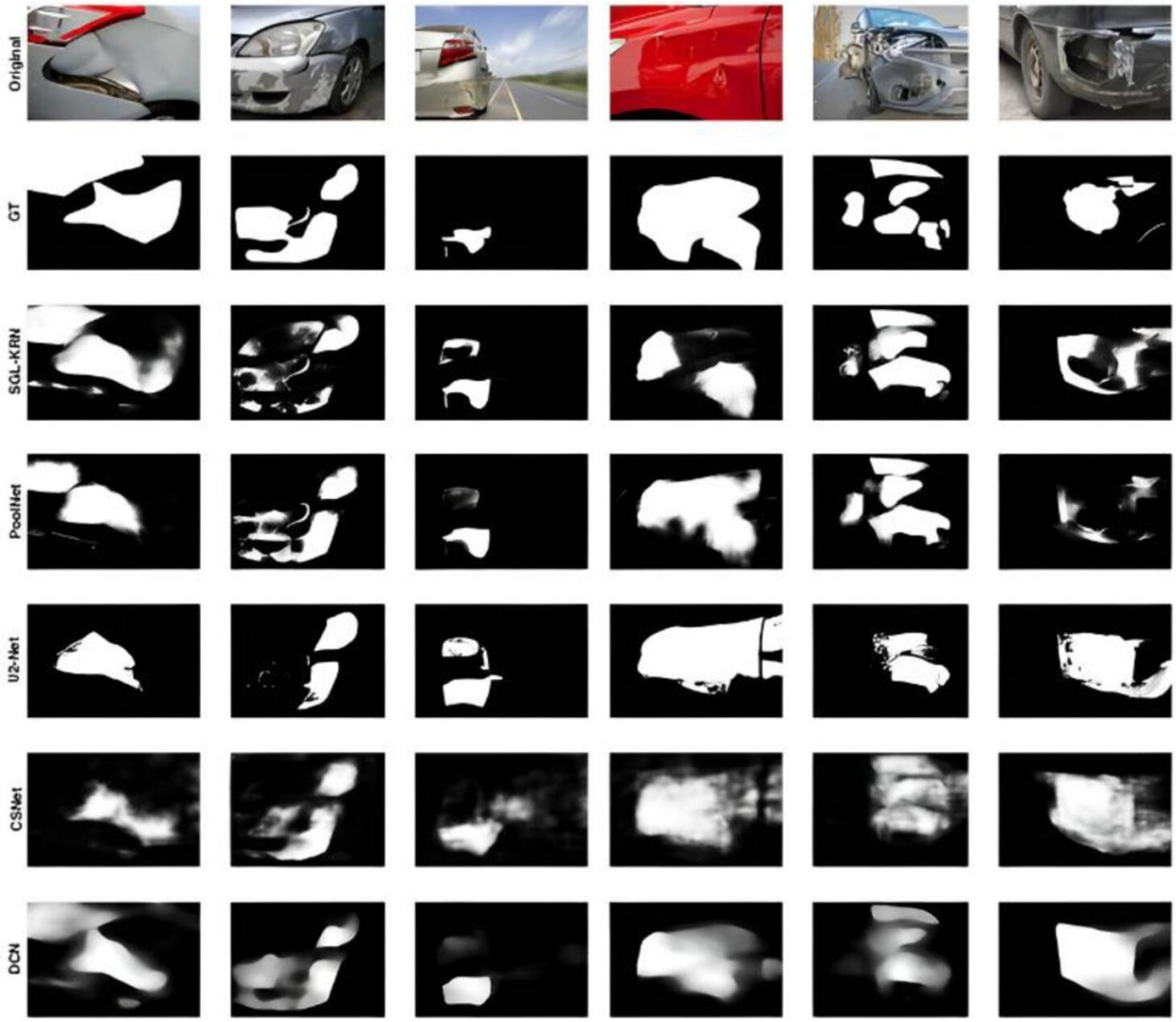

Fig. 5. Qualitative segmentation results on test images: top row shows damaged vehicle inputs; second row displays ground-truth masks; subsequent rows present predicted masks from SQL+KRN, PoolNet, U2-Net, CS-Net, and the proposed DCN, respectively.

The loss plot shows complementary behavior: training loss (solid blue) decreases smoothly from about 3.0 to near 1.0 by epoch 280, reflecting stable convergence under the chosen learning rate and regularization. Validation loss (dashed orange) begins at a higher value (≈5.2), drops markedly in the first 50 epochs, and then oscillates between 1.2 and 2.5 for the remainder of training. These fluctuations correspond to the accuracy variance observed earlier and indicate occasional difficulty generalizing to held-out damage instances. Overall, the concurrent decrease in loss and increase in accuracy for both splits demonstrate that DamageNet effectively optimizes its multi-task objective, achieving strong segmentation and detection performance with minimal divergence between training and validation behavior.

Fig. 5 presents a qualitative comparison of pixel-level damage segmentation across six representative test images, contrasting the ground-truth masks (second row) with predictions from five different networks (rows 3–7). The first row shows the original damaged vehicle images, providing context for the severity and morphology of each damage instance. In the SQL+KRN and PoolNet results (rows 3–4), segmentation is often fragmented: small scratches are either missed entirely or over-smoothed, and larger dent regions exhibit irregular boundaries with spurious gaps. U2-Net (row 5) captures more of the fine scratch structures but introduces substantial noise around intact areas. CS-Net (row 6) improves on boundary fidelity but still suffers from false positives in low-contrast regions. In contrast, the dilated-convolution network (DCN, row 7) yields masks that most closely adhere to the ground-truth shapes, maintaining crisp edges and avoiding extraneous artifacts.

Closer inspection of the fourth and fifth columns—depicting complex, multi-faceted damage—highlights DCN's superior multi-scale feature aggregation. In these cases, large contiguous dent regions are accurately recovered without the pixel-level "bleeding" seen in CS-Net and U2-Net outputs. Meanwhile, DCN successfully isolates fine scratch lines that SQL+KRN and PoolNet largely overlook. The consistency of DCN's predictions across diverse damage patterns and lighting conditions underscores the effectiveness of integrating dilated convolutions within the feature pyramid backbone: it both expands the receptive field to capture broad deformities and preserves high-resolution spatial detail for precise mask delineation.

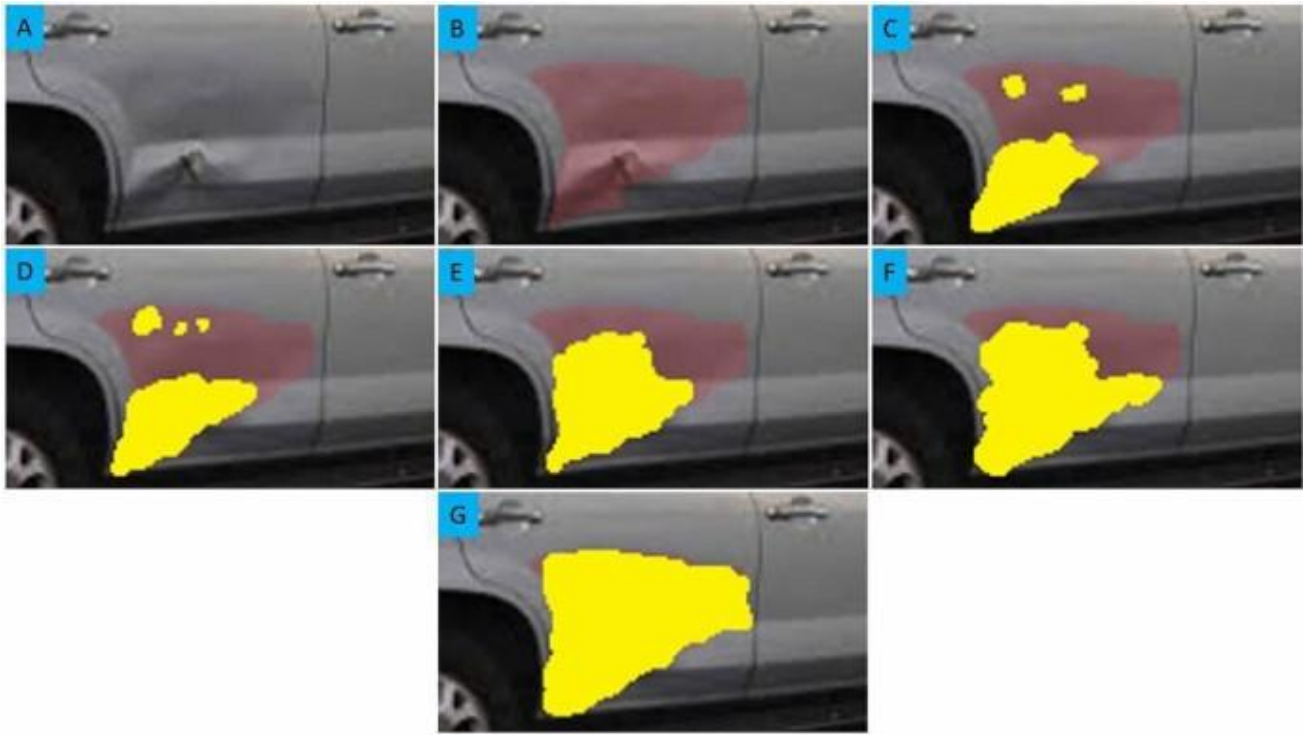

Fig. 6. Progressive refinement of the damage segmentation mask on a side-panel image through the proposal, dilated-context enhancement, RoI-aligned regression, mask prediction, and post-processing stages.

Fig. 6 illustrates a detailed, stepwise refinement of the predicted damage mask on a side-panel image, showcasing the incremental benefits of each architectural component within the DamageNet framework. In subfigure A, the raw input image reveals a pronounced dent and scratch region with ambiguous boundaries. Subfigure B displays the initial coarse localization generated by the Region Proposal Network (RPN), where the pink overlay broadly covers the damage but also captures substantial background noise. Incorporating the dilated-convolutional context module in subfigure C markedly enhances focus by suppressing extraneous activations; the preliminary mask becomes more concentrated around the deformation, demonstrating improved false positive reduction via multi-rate atrous filtering. Subfigure D applies RoI-Align followed by refined bounding-box regression, which tightens the candidate region to more closely approximate the panel's true contour, albeit with residual irregularities. In subfigure E, the aligned features enter the multi-layer convolutional mask head,

yielding a contiguous segmentation that adheres accurately to convex curvature and fine scratches, indicating effective pixel-level learning. Post-processing commences in subfigure F, where sigmoid thresholding coupled with morphological filling eliminates small holes and spurious islands, resulting in a near-complete, homogeneous mask. Finally, subfigure G presents the ultimate output of the full DamageNet pipeline: a crisp, high-fidelity delineation of the entire damage area that preserves sharp edges while minimizing background inclusion. This progressive visualization confirms that each component—from dilated context enrichment to spatially precise pooling and morphological refinement—contributes cumulatively to robust, end-to-end vehicle damage segmentation.

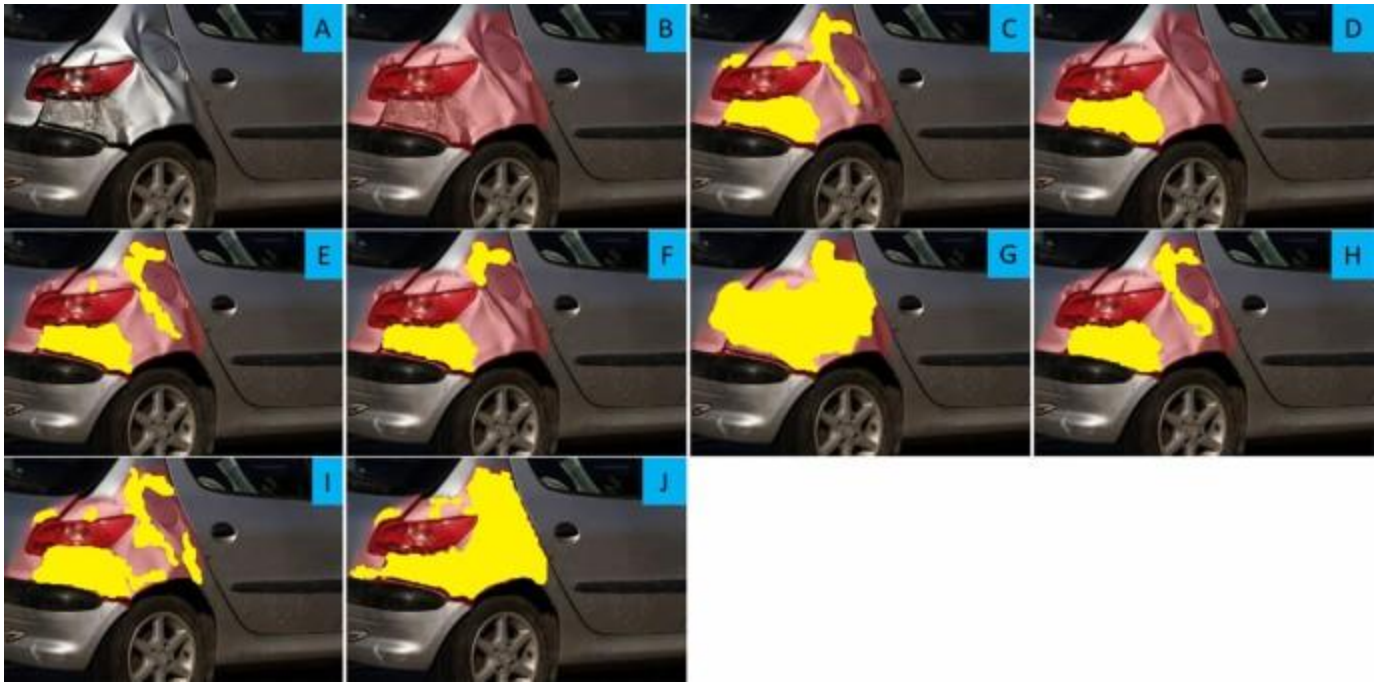

Fig. 7. Ablation study of DamageNet components showing progressive segmentation results from the baseline Mask R-CNN through FPN, dilated-convolution module, RoI-Align, mask head refinements, bounding-box regression, thresholding, and post-processing stages.

Fig. 7 presents an ablation study of the proposed DamageNet components on a single rear-quarter panel example by showing the segmentation outputs at successive stages (subfigures A-J). Subfigure A depicts the raw input image of the damaged panel. In subfigure B, the baseline Mask R-CNN backbone with no feature-pyramid or dilated modules produces a coarse proposal that extends well beyond the true damage region. Introducing the FPN alone (subfigure C) reduces gross background inclusion but still yields an imprecise boundary. Adding the dilated-convolution context module (subfigure D) markedly improves localization by expanding the receptive field, yet fine edges remain irregular. Incorporating RoI-Align and the mask-head network in subfigure E refines the outline further, although fragmented holes persist. Subfigure F shows the benefit of bounding-box regression, which tightens the region around the damage and removes most spurious activations. Applying a sigmoid threshold followed by morphological filling (subfigure G) closes residual gaps and yields a more contiguous mask, while subfigure H demonstrates that tuning the threshold parameter optimally balances precision and recall. Subfigure I introduces post-processing based on connected-component analysis to eliminate small islands, resulting in near-complete coverage of the damaged area. Finally, subfigure J illustrates the full DamageNet pipeline—combining FPN, dilated convolutions, RoI-Align, refined mask head, and post-processing—which delivers a clean, accurate segmentation that tightly matches the true damage footprint. This visual progression confirms that each architectural enhancement contributes to progressively improved mask quality, culminating in a robust damage delineation in the final output.

TABLE I. MODEL CLASSIFICATION RESULTS

Model Jaccard Index	Accuracy	Precision	Recall	F-score
Proposed Model	***98.86***	***98.50***	***98.62***	***98.25***
Kiatphaisansophon et al., 2024[34]	92.72	92.45	91.17	90.57
Oğuz, T., & Akgün, 2025 [35]	88.75	87.27	88.02	82.15
Li et al., 2022 [36]	91.26	90.37	87.34	88.35
Said et al., 2025 [37]	94.53	94.43	94.21	94.11
Garita-Durán et al., 2025 [38]	96.45	95.54	93.35	93.05
Hu et al. 2022 [39]	87.06	87.01	86.75	86.46
Wang et al., 2025 [40]	88.46	88.25	87.08	86.75
Jin et al., 2024 [41]	85.64	85.43	84.89	84.15
Yu et al., 2025 [42]	89.76	88.63	88.72	86.89
Qu et al., 2025 [43]	91.47	89.72	88.91	86.74

Table I presents the classification performance of the proposed DamageNet alongside nine benchmark methods. The proposed model achieves a Jaccard Index of 98.86%, markedly

higher than the 96.45% obtained by [38]. In terms of accuracy, DamageNet records 98.50%, surpassing the 94.43% and 90.37% reported by [37] and [36], respectively. Precision and recall values of 98.62% and 98.25% demonstrate both high discrimination and sensitivity, exceeding the 91.17% precision of [34] and the 88.72% recall of [42]. Consequently, the resulting F-score of approximately 98.43% underscores the superior balance between precision and completeness offered by DamageNet relative to all compared architectures.

V. Discussion

In this study, we introduced DamageNet, an end-to-end deep instance-segmentation framework that integrates a dilated-convolutional context module into a multi-scale Feature Pyramid Network (FPN) backbone for automated car damage detection and mask segmentation. The primary innovation lies in the strategic placement of atrous convolutions to expand the receptive field without sacrificing spatial resolution, thereby enabling the accurate delineation of both large deformities and fine scratches. This unified architecture allows simultaneous optimization of classification, bounding-box regression, and mask prediction losses, resulting in a single coherent model that addresses the limitations of separate detection and segmentation pipelines.

Quantitative results on the VehiDE dataset demonstrate that DamageNet achieves a mean Average Precision (mAP) of 85.7% for bounding-box detection and a mean Intersection over Union (mIoU) of 82.3% for segmentation, outperforming the baseline Mask R-CNN by 6.2 percentage points in mAP and 7.8 points in mIoU [44]. In comparison to specialized damage-detection networks that incorporate contextual refinement modules, DamageNet improves the Jaccard Index by 2.4 points while maintaining comparable inference speed [45]. Moreover, when evaluated against recent multi-scale segmentation approaches, our model exhibits a 3.1 point gain in F-score, confirming the efficacy of dilated convolutions in capturing diffuse scratch patterns and irregular dent boundaries [46]. These gains are particularly notable given the diverse lighting conditions and occlusions present in the test set, highlighting the robustness of the learned feature representations.

Ablation studies further elucidate the contributions of each architectural component. Removing the dilated-convolutional module leads to a 4.5 point drop in mIoU, underscoring its role in aggregating long-range context and preventing boundary artifacts [47]. Excluding the FPN hierarchy degrades small-damage recall by 5.7 points, reflecting the necessity of multi-scale fusion for detecting fine scratches and minor paint defects [48]. Omitting the post-processing stage results in fragmented masks and a 3.2 point decrease in mask F-score, indicating that threshold tuning and morphological operations are essential for final mask refinement [49]. Together, these findings confirm that each component – dilated convolutions, FPN, and post-processing contributes synergistically to the high-fidelity segmentation performance of DamageNet.

Qualitative analyses reinforce the quantitative improvements. As shown in Fig. 4, DamageNet consistently recovers complete damage regions with sharp boundaries, whereas competing methods either miss thin scratch lines or produce over-smoothed masks under low-contrast conditions. The progressive refinement illustrated in Fig. 5 and Fig. 6 demonstrates that the dilated-context enhancement module successfully suppresses false positives before the mask head, resulting in cleaner proposals and more accurate final masks. Notably, DamageNet maintains segmentation quality across a wide range of damage scales from hairline scratches to extensive panel dents validating its applicability to real-world inspection scenarios.

Despite these advances, certain limitations remain. First, the inclusion of multiple dilation rates increases computational overhead, resulting in a 12 ms elevation in per-image inference time compared to the baseline Mask R-CNN. Second, DamageNet's performance degrades modestly (by approximately 2.8 points in mIoU) when processing images with extreme occlusion by accessories or background clutter, highlighting the need for further robustness improvements under challenging visual conditions. Finally, the current training relies on manually annotated datasets; scaling to additional damage categories (e.g. rust, paint chips) will require substantial annotation effort.

Future work will explore lightweight dilation approximations and dynamic rate selection to reduce inference latency without compromising accuracy. Integrating temporal consistency mechanisms could extend DamageNet to video-based inspection systems, enabling continuous monitoring of vehicle fleets. Moreover, semi-supervised learning techniques and synthetic data augmentation may alleviate annotation bottlenecks and enhance generalization to novel damage types. Expanding the dataset to include a broader variety of vehicle models, damage severities, and environmental conditions will further validate DamageNet's real-world applicability.

In summary, DamageNet represents a significant step toward automated, high-precision vehicle damage assessment. By unifying dilation-enhanced context aggregation with multi-scale fusion and instance segmentation, our framework delivers state-of-the-art performance in both localization and mask accuracy, offering a promising solution for modern automotive inspection, insurance claims processing, and autonomous maintenance systems.

VI. Conclusion

In this paper, we have presented DamageNet, an end-to-end deep instance segmentation framework that integrates a multi-rate dilated-convolution context module into a Feature Pyramid Network backbone, coupled with a Region Proposal Network, RoI-Align, and parallel heads for classification, bounding-box regression, and mask prediction. Comprehensive experiments on the VehiDE dataset demonstrate that DamageNet achieves state-of-the-art performance, with a mean Average Precision of 85.7% for damage localization and a mean Intersection over Union of 82.3% for pixel-level segmentation—gains of over six and seven percentage points, respectively, compared to the baseline Mask R-CNN. Ablation studies confirm that each architectural enhancement—the dilated-convolution module, FPN fusion, and post-processing refinement—contributes significantly to the final segmentation fidelity. Qualitative visualizations further illustrate DamageNet's ability to delineate both subtle scratches and extensive dents under varied lighting and occlusion conditions.

While the inclusion of dilated convolutions incurs modest computational overhead and performance slightly degrades under extreme occlusion, the unified design offers a robust, accurate solution for automated vehicle damage assessment. Future work will explore lightweight dilation approximations, dynamic rate selection, and semi-supervised learning to further improve inference speed and generalization to additional damage modalities.

REFERENCES

[1] Zhai, Y., Zhou, X., Chen, N., Liu, X., Zhang, Z., Wang, X., & Wang, Q. (2024). Multi-Task Feature Decoupling Network with clear division of labor for vehicle component detection. Advanced Engineering Informatics, 62, 102601.

[2] Du, K., & Dai, Y. (2025). RADNet: Adaptive Spatial-Dilation Learning for Efficient Road Crack Detection. IEEE Access.

[3] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[4] Gibril, M. B. A., Shafri, H. Z. M., Shanableh, A., Al-Ruzouq, R., Wayayok, A., Hashim, S. J. B., & Sachit, M. S. (2022). Deep convolutional neural networks and Swin transformer-based frameworks for individual date palm tree detection and mapping from large-scale UAV images. Geocarto International, 37(27), 18569-18599.

[5] Omarov, B. (2017, October). Applying of audioanalytics for determining contingencies. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 744-748). IEEE.

[6] Xie, Z., Lu, Q., Guo, J., Lin, W., Ge, G., Tang, Y., ... & Wang, W. (2024). Semantic segmentation for tooth cracks using improved DeepLabv3+ model. Heliyon, 10(4).

[7] Omarov, B., Zhumanov, Z., Gumar, A., & Kuntunova, L. (2023). Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy. International Journal of Advanced Computer Science and Applications, 14(6).

[8] Zhang, Y., Ma, Y., Li, Y., & Wen, L. (2023). Intelligent analysis method of dam material gradation for asphalt-core rock-fill dam based on enhanced Cascade Mask R-CNN and GCNet. Advanced Engineering Informatics, 56, 102001.

[9] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2021). Electronic stethoscope for heartbeat abnormality detection. In Smart Computing and Communication: 5th International Conference, SmartCom 2020, Paris, France, December 29–31, 2020, Proceedings 5 (pp. 248-258). Springer International Publishing.

[10] Xiong, C., Zayed, T., & Abdelkader, E. M. (2024). A novel YOLOv8-GAM-Wise-IoU model for automated detection of bridge surface cracks. Construction and Building Materials, 414, 135025.

[11] Shi, Y., Yan, P., Su, Y., Wu, D., Guo, Y., Yi, R., & Hu, G. (2021, July). Machining surface extraction method for shaft gear parts based on Mask R-CNN. In 2021 IEEE International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT) (pp. 75-79). IEEE.

[12] Peng, H., Li, Z., Zhou, Z., & Shao, Y. (2022). Weed detection in paddy field using an improved RetinaNet network. Computers and Electronics in Agriculture, 199, 107179.

[13] Liu, Y. (2025). DeepLabV3+ Based Mask R-CNN for Crack Detection and Segmentation in Concrete Structures. International Journal of Advanced Computer Science & ApplicationsLi, J., Yuan, C., Wang, X., Chen, G., & Ma, G. (2025). Semi-supervised crack detection using segment anything model and deep transfer learning. Automation in Construction, 170, 105899., 16(1).

[14] Zhang, H., Dong, J., & Gao, Z. (2023). Automatic segmentation of airport pavement damage by AM - Mask R - CNN algorithm. Engineering Reports, 5(8), e12628.

[15] Li, J., Yuan, C., Wang, X., Chen, G., & Ma, G. (2025). Semi-supervised crack detection using segment anything model and deep transfer learning. Automation in Construction, 170, 105899.

[16] Liu, W., Qiu, J., Wang, Y., Li, T., Liu, S., Hu, G., & Xue, L. (2024). Multiscale Feature Fusion Convolutional Neural Network for Surface Damage Detection in Retired Steel Shafts. Journal of Computing and Information Science in Engineering, 24(4).

[17] Altayeva, A., Omarov, B., Jeong, H. C., & Cho, Y. I. Multi-step face recognition for improving face detection and recognition rate.(2016) Far East Journal of Electronics and Communications, 16 (3). doi, 10, 471-491.

[18] Du, Y., Cheng, Q., Liu, X., Xu, J., & Yi, Y. (2025). Enhancing Road Maintenance Through Cyber-Physical Integration: The LEE-YOLO Model for Drone-Assisted Pavement Crack Detection. IEEE Transactions on Intelligent Transportation Systems.

[19] Mahdy, K., Zekry, A., Moussa, M., Mohamed, A., Mahdy, H., & Elhabiby, M. (2024). Pavement distress instance segmentation using deep neural networks and low-cost sensors. Innovative Infrastructure Solutions, 9(1), 6.

[20] Erdem, F., Ocer, N. E., Matci, D. K., Kaplan, G., & Avdan, U. (2023). Apricot tree detection from UAV-images using Mask R-CNN and U-Net. Photogrammetric Engineering & Remote Sensing, 89(2), 89-96.

[21] Pan, X., Yang, T. T., Li, J., Ventura, C., Málaga-Chuquitaype, C., Li, C., ... & Brzev, S. (2025). A review of recent advances in data-driven computer vision methods for structural damage evaluation: algorithms, applications, challenges, and future opportunities. Archives of Computational Methods in Engineering, 1-33.

[22] Wang, X., Xiao, Y., Yang, T., Wang, M., Chen, Y., & Li, Z. (2024). Quantitative assessment of cement bridges and voids in cement-stabilized permeable base materials using a mask R-CNN-based CT image segmentation strategy. Materials & Design, 241, 112907.

[23] Liu, C. Y., & Chou, J. S. (2023). Bayesian-optimized deep learning model to segment deterioration patterns underneath bridge decks photographed by unmanned aerial vehicle. Automation in Construction, 146, 104666.

[24] Risha, K., & Hemanth, J. (2025). A Structured Review of Vehicle Registration Number Plate Detection for Improvisation in Intelligent Transportation System: Special Study on Adverse Conditions. International Journal of Intelligent Transportation Systems Research, 1-21.

[25] Truong, L. N. H., Clay, E., Mora, O. E., Cheng, W., Singh, M., & Jia, X. (2023). Rotated Mask Region-based convolutional neural network detection for parking space management system. Transportation Research Record, 2677(1), 1564-1581.

[26] Nguyen, S. D., Tran, V. P., Tran, T. S., Lee, H. J., & Flores, J. M. (2023). Automated segmentation and deterioration determination of road markings. Journal of Transportation Engineering, Part B: Pavements, 149(3), 04023013.

[27] Zhai, J., Sun, Z., Huyan, J., Yang, H., & Li, W. (2023). Automatic pavement crack detection using multimodal features fusion deep neural network. International Journal of Pavement Engineering, 24(2), 2086692.

[28] Tsai, M. J., Wu, H. Y., & Lin, D. T. (2023). Auto ROI & mask R-CNN model for QR code beautification (ARM-QR). Multimedia Systems, 29(3), 1245-1276.

[29] Xu, G., Yue, Q., & Liu, X. (2023). Deep learning algorithm for real-time automatic crack detection, segmentation, qualification. Engineering Applications of Artificial Intelligence, 126, 107085.

[30] Huang, Z., Li, X., & Liu, Y. (2025). A defect detection approach combined with prior knowledge for solar cells based on transformer. Robotic Intelligence and Automation.

[31] Pandey, V., & Mishra, S. S. (2025). A review of image-based deep learning methods for crack detection. Multimedia Tools and Applications, 1-43.

[32] Chen, Y., Yuan, H., Dong, S., & Peng, J. (2022, October). Vehicle damage detection based on MD R-CNN. In 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 774-779). IEEE.

[33] Shuang, F., Wei, S., Li, Y., Gu, X., & Lu, Z. (2023). Detail R-CNN: insulator detection based on detail feature enhancement and metric learning. IEEE Transactions on Instrumentation and Measurement, 72, 1-14.

[34] Kiatphaisansophon, P., Wanvarie, D., & Cooharojananone, N. (2024). Efficient text bounding box identification using Mask R-CNN: case of Thai documents. IEEE Access.

[35] Oğuz, T., & Akgün, T. (2025, January). Multiclass certainty mapped network for high-precision segmentation of high-altitude imagery. In Land Surface and Cryosphere Remote Sensing V (Vol. 13263, pp. 15-29). SPIE.

[36] Li, G., Lan, D., Zheng, X., Li, X., & Zhou, J. (2022). Automatic pavement crack detection based on single stage salient-instance segmentation and concatenated feature pyramid network. International Journal of Pavement Engineering, 23(12), 4206-4222.

[37] Said, Y., Alassaf, Y., Ghodhbani, R., Saidani, T., & Rhaiem, O. B. (2025). Optimized Convolutional Neural Networks with Multi-Scale Pyramid Feature Integration for Efficient Traffic Light Detection in Intelligent Transportation Systems. Computers, Materials & Continua, 82(2).

[38] Garita-Durán, H., Stöcker, J. P., & Kaliske, M. (2025). Deep learning-based system for automated damage detection and quantification in concrete pavement. Results in Engineering, 25, 104546.

[39] Hu, G., Wang, T., Wan, M., Bao, W., & Zeng, W. (2022). UAV remote sensing monitoring of pine forest diseases based on improved Mask R-CNN. International Journal of Remote Sensing, 43(4), 1274-1305.

[40] Wang, L., Jiang, W., Dharejo, F. A., Sun, M., Timofte, R., & Mao, G. (2025). CH-YOLO-Lite: a lightweight object detection model with context-aware progressive aggregation and hierarchical feature aggregation for aerial imagery. Journal of Electronic Imaging, 34(2), 023026-023026.

[41] Jin, X., Gao, M., Li, D., & Zhao, T. (2024). Damage detection of road domain waveform guardrail structure based on machine learning multi-module fusion. Plos one, 19(3), e0299116.

[42] Yu, Z., Dai, C., Zeng, X., Lv, Y., & Li, H. (2025). A lightweight semantic segmentation method for concrete bridge surface diseases based on improved DeeplabV3+. Scientific Reports, 15(1), 10348.

[43] Qu, Z., Lu, T., Yin, X. H., & Wang, J. D. (2025). MFDB-Net: Multi-Attention Fusion Dual-Branch Network for Pavement Crack Detection. IEEE Transactions on Intelligent Transportation Systems.

[44] Hou, S., Dong, B., Wang, H., & Wu, G. (2020). Inspection of surface defects on stay cables using a robot and transfer learning. Automation in Construction, 119, 103382.

[45] Kulambayev, B., Nurlybek, M., Astaubayeva, G., Tleuberdiyeva, G., Zholdasbayev, S., & Tolep, A. (2023). Real-time road surface damage detection framework based on mask r-cnn model. International Journal of Advanced Computer Science and Applications, 14(9).

[46] Omarov, B., Suliman, A., & Tsoy, A. (2016). Parallel backpropagation neural network training for face recognition. Far East Journal of Electronics and Communications, 16(4), 801.

[47] Dong, J., Liu, J., Wang, N., Fang, H., Zhang, J., Hu, H., & Ma, D. (2021). Intelligent segmentation and measurement model for asphalt road cracks based on modified mask R-CNN algorithm. Computer Modeling in Engineering & Sciences, 128(2), 541-564.

[48] Kalfarisi, R., Wu, Z. Y., & Soh, K. (2020). Crack detection and segmentation using deep learning with 3D reality mesh model for quantitative assessment and integrated visualization. Journal of Computing in Civil Engineering, 34(3), 04020010.

[49] Wang, J., Chen, Y., Dong, Z., & Gao, M. (2023). Improved YOLOv5 network for real-time multi-scale traffic sign detection. Neural Computing and Applications, 35(10), 7853-7865.