Multi-Task Vehicle Surface Damage Analysis Model Based on YOLOv8

Xu Tan, Ji Zhao*

***Abstract*—The rapid evolution of intelligent automotive systems has driven the urgent need for advanced damage assessment methodologies, revealing critical limitations in conventional visual inspection techniques. This study posits a multi-task learning architecture for vehicular surface damage quantification through synergistic integration of instance segmentation and monocular depth estimation. Three key technical innovations are incorporated: 1) A Heterogeneous Feature Single-Phase Booster (HFSPB) module utilizing RepViT layers to optimize backbone feature discriminability; 2) A Channel-wise Cross Fusion Block facilitating adaptive multi-scale feature amalgamation with intrinsic noise attenuation; 3) A depth estimation head implementing Depth Interval Attraction Refinement for geometrically consistent surface reconstruction. Benchmark evaluations revealed a threefold acceleration in inference speed relative to conventional approaches, coupled with robust metric performance (δ_1: 91.9%, δ_2: 99.4%, δ_3: 99.8%, REL=0.089, RMSE=0.312). This integrated framework establishes a computationally efficient paradigm for multimodal damage characterization, providing critical insights for autonomous vehicle maintenance systems.**

***Index Terms*—Automotive Damage Analysis, Monocular Depth Estimation, YOLOv8, Attention Mechanism.**

I. Introduction

THE proliferation of intelligent automotive systems has fundamentally reshaped the vehicle diagnostics paradigm, as the growing vehicle population necessitates enhanced diagnostic throughput. Conventional damage assessment protocols are typically performed through manual inspections by mechanics, involving sequential evaluations of structural components (e.g., bumpers, hoods, door panels, and load-bearing pillars) to identify surface defects such as scratches, dents, and perforations. This experience-driven methodology, while historically effective, increasingly fails to meet the precision and scalability requirements of next-generation intelligent transportation ecosystems.

Contemporary deep learning frameworks have revolutionized automotive defect identification, superseding conventional approaches reliant on hand-engineered feature descriptors. Sophisticated computational tasks, including object localization and pixel-wise semantic segmentation, are now achievable through end-to-end architectures that precisely delineate damage morphology and spatial distribution. However, prevailing architectures are predominantly characterized by network depth escalation and feature extractor optimization, while underrepresenting task-specific characteristics inherent to vehicular surface defect analysis. Furthermore, systematic investigation of composite damage patterns remains conspicuously absent in existing implementations.

The paradigm shift towards intelligent transportation ecosystems necessitates next-generation vehicle damage analysis systems to integrate multifunctional detection capabilities, notably object localization and instance segmentation, while generating geometrically annotated metadata beyond basic defect identification. Critically, such systems should incorporate standardized interfaces for multimodal damage interpretation, enabling coordinated execution of integrated workflows encompassing defect detection, component-level segmentation, and structural integrity quantification.

This study proposes a novel framework integrating YOLOv8-based monocular depth estimation with vehicle damage assessment, extending our prior architecture. Feature extraction capabilities were enhanced through a Heterogeneous Feature Single-Phase Booster (HFSPB) module, implemented through RepViT operators that optimize multi-dimensional feature discrimination. The transformed multi-scale data were subsequently encoded into hierarchical embeddings, systematically serialized as Transformer tokens, and processed through a Channel-wise Cross Fusion Block (CCFB) to establish cross-modal feature correlations. The architecture culminates in a depth estimation head incorporating deep interval attraction refinement, achieving enhanced computational efficiency while maintaining accuracy degradation compared to conventional approaches. Notably, this research pioneers the integration of instance segmentation and monocular depth estimation within a unified multi-task paradigm. The synergistic combination of complementary 3D depth data and 2D segmentation matrices enables systematic damage characterization through three-dimensional geometric reconstruction and surface topology analysis, particularly critical for assessing collision-induced deformations.

II. Related Work

The automotive damage assessment field has undergone a fundamental methodological evolution, progressing from conventional manual feature engineering to data-driven deep

Manuscript received Nov 18, 2024; revised Apr 23, 2025.

This work was supported by the Natural Science Foundation of China (No.62272093) and by Liaoning Key Laboratory of the Internet of Things Application Technology on Intelligent Construction.

Xu Tan is a graduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 1304834230@qq.com).

Ji Zhao is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 319973500069@ustl.edu.cn).

learning architectures. This transformation has enabled the development of advanced computer vision tasks including object detection and semantic segmentation, which systematically identify structural deformations and surface anomalies through pixel-level pattern recognition.

A. Deep Learning Methods

Contemporary automotive damage assessment predominantly employs deep learning architectures, with object detection and semantic segmentation emerging as principal modalities for vehicular image analysis. Jiao[1] enhanced the Faster R-CNN framework through post-RoI-pooling integration of an Online Hard Example Minimization algorithm, enabling systematic identification of tire defects in X-ray radiographic data. Zhu et al.[2] implemented a dual-network architecture combining Faster R-CNN with generative adversarial networks for image quality augmentation, while upgrading the backbone from VGG16 to ResNet101 with Feature Pyramid Network integration to enhance semantic feature representation. Notably, Sun et al.[3] developed a non-local U-Net variant incorporating spatiotemporal dependency modules for bearing defect segmentation through cross-dimensional feature correlation. Parallel advancements by Tang et al.[4] leveraged HRNet-DeepLabv3+ hybrid architectures with multi-scale atrous spatial pyramid pooling, establishing hierarchical feature fusion mechanisms for wheel defect analysis.

The YOLO architecture series has driven significant advancements in real-time automotive defect detection frameworks. Zhang et al.[5] proposed a YOLOv3-SPP variant incorporating spatial pyramid pooling (SPP) modules within the backbone network to synthesize multi-scale feature representations, enhancing recognition performance for nine distinct wheel region defect categories. In parallel developments, Lv[6] optimized the YOLOv5 architecture through integration of an Adaptive Structural Feature Fusion (ASFF) mechanism coupled with Convolutional Block Attention Module (CBAM) components, while implementing a ShuffleNetv2-based lightweight backbone, achieving efficient detection of glass surface anomalies including bubble formations and scratch patterns.

B. EIS-YOLO

Building upon our prior work[7], an enhanced instance segmentation framework (EIS-YOLO) was developed for automotive damage assessment, leveraging the YOLOv8 architecture as diagrammed in Figure 1. The core innovation involved structural reconfiguration of the backbone network through replacement of conventional C2f modules with multi-scale Channel-Reduction Dense Block (CRDB) units, achieving 20.15% parameter reduction while enhancing multi-level feature fusion efficacy on the CarDD benchmark[8]. Notably, a High-Resolution Feature Pyramid Network (HRFPN) was implemented to preserve spatial granularity through persistent high-resolution branches, synergistically integrated with Attention Feature Fusion (AFF) and Bidirectional Attention Module (BiAM) components to strengthen cross-scale information propagation. The enhanced feature pyramid (E-FPN) architecture further optimized inter-layer connectivity through streamlined skip connections, complemented by a dedicated micro-damage detection head specifically engineered for small target recognition and edge delineation. Quantitative evaluation demonstrated respective 4.4% (P_B) and 6.6% (P_M) accuracy improvements over baseline models.

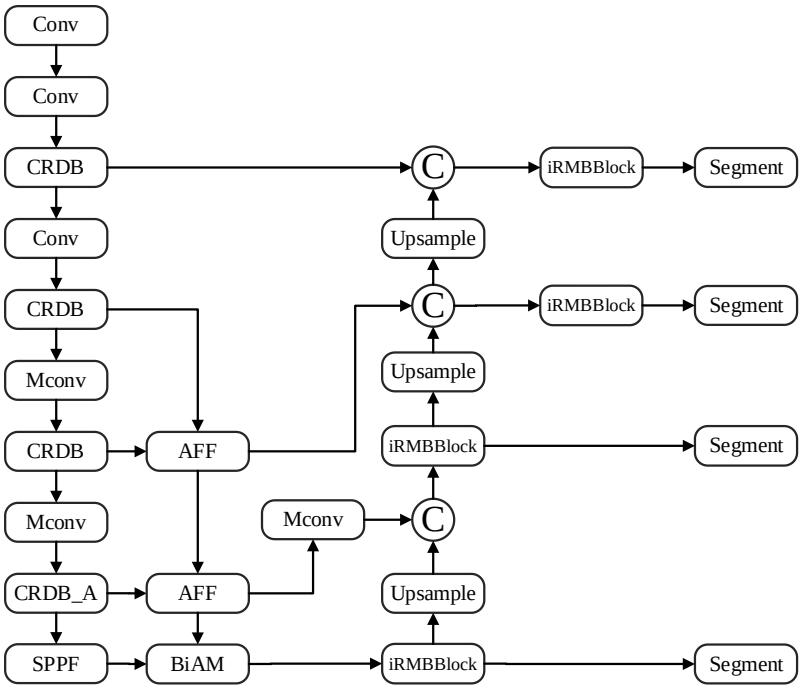

Fig. 1. EIS-YOLO Structure Diagram

The acquisition of three-dimensional structural data is conventionally achieved through photogrammetric modalities including LiDAR, structured light illumination, Time-of-Flight(ToF) sensors, and structured stereo reconstruction systems. However, monocular depth estimation has emerged as the preferred methodology when considering cost-efficiency and operational versatility, despite constituting an ill-posed inverse problem where three-dimensional geometry must be reconstructed from two-dimensional projective measurements with inherent depth information loss.

Traditional computational geometry approaches relied on perspective-n-point algorithms and hand-engineered feature descriptors, attempting depth recovery through geometric priors such as vanishing point convergence and projective size variation. Parallel methodologies exploited manual texture analysis using gradient operators, edge detection filters, and photometric stereo techniques to infer surface orientation. Nevertheless, these heuristic methods demonstrated limited reliability due to their dependence on scene-specific texture patterns and predefined object shape assumptions.

The paradigm has fundamentally shifted with the advent of deep learning architectures, transitioning from manual feature engineering to data-driven end-to-end frameworks. Contemporary convolutional neural networks and vision transformers now enable robust depth prediction through hierarchical feature abstraction, achieving superior cross-domain generalization compared to conventional computer vision techniques while maintaining computational tractability.

III. METHODS

Traditional monocular depth estimation employs static depth bin discretization, formulating prediction as classification task through fixed interval quantization. This approach suffers from error propagation due to inappropriate bin configuration sensitivity.

Our architecture introduces three key innovations: First, a dynamic bin allocation mechanism adaptively optimizes depth intervals during training, extending ZoeDepth's [9]attraction strategy. Second, the High-Frequency Spatial Pyramid Block (HFSPB) mitigates feature degradation through multi-scale spatial recalibration, addressing inherent signal attenuation in depth tasks. Third, the Cross-Channel Fusion Block (CCFB) integrates channel attention with dilated convolutions to establish cross-scale dependencies while preserving spatial coherence.

The depth estimation head synergistically combines these components with gated skip connections from the backbone network, forming an integrated multi-task framework for vehicular damage analysis as detailed in Figure 2. Subsequent sections delineate each module's operational mechanics and topological integration.

A. Depth Interval Attraction Refinement Strategy

The primary objective of the Metric Bin Module is to categorize depth values into a defined set of depth bins. Each bin is associated with a defined depth range, and the model forecasts a pixel's depth by evaluating the probability distribution over these bins. This approach enhances prediction accuracy and improves the management of uncertainty in depth estimation. The module initiates by establishing the boundaries of depth bins, with the depth range from D_{min} to D_{max} divided into k equally spaced bins. The limits of each bin can therefore be established as:

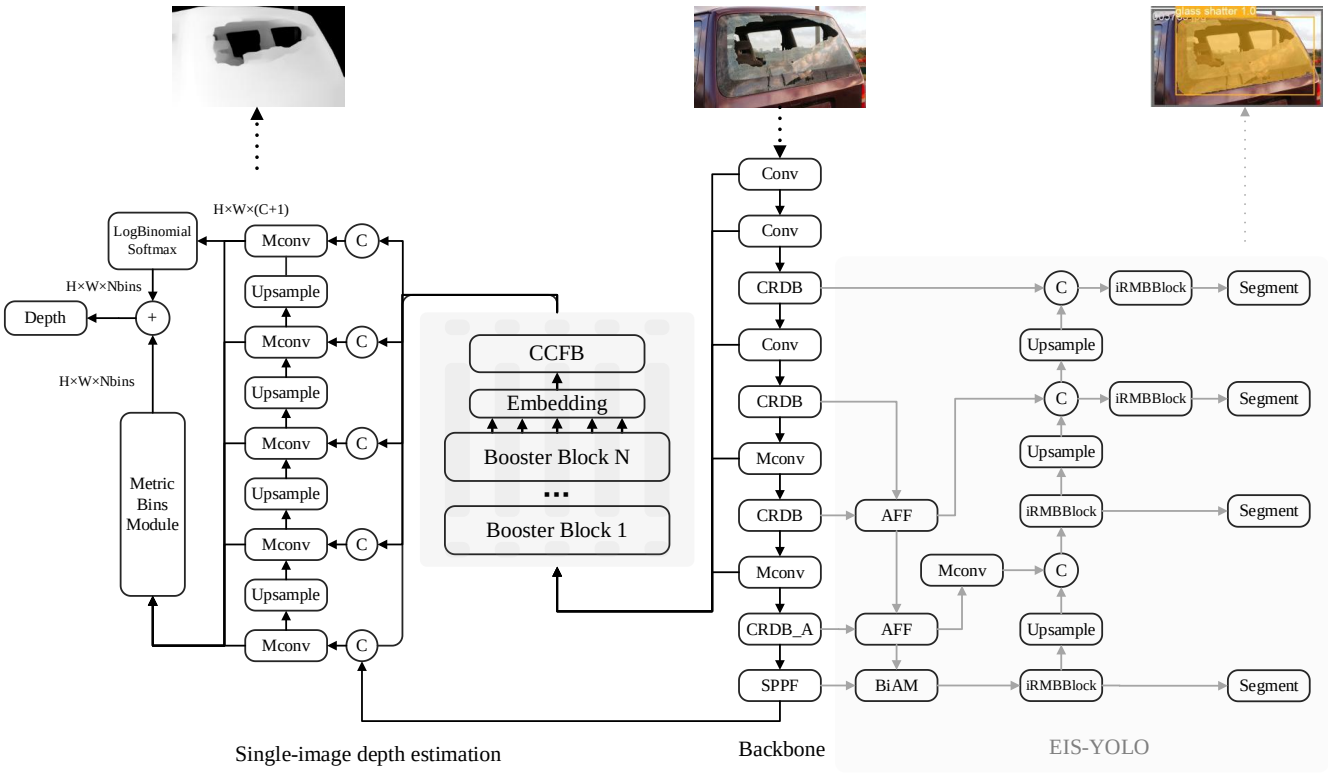

Fig. 2. Multi-task Large Model for Vehicle Exterior Damage

$$b_k = D_{min} + k \cdot \Delta D, k = 0,1,2,\ldots,K \quad (1)$$

$$\Delta D = \frac{D_{max} - D_{min}}{K} \quad (2)$$

The equation 2 defines the depth span of each bin, with the depth range for the i-th depth bin expressed as $[b_i, b_{i+1}]$. For each pixel in the input image, the model predicts a probability distribution for each depth bin via the Metric Bin Module. Let the probability distribution for this pixel point be represented as $P=[p_1, p_2, \ldots, p_K]$, where the Equation 3 must hold true.

$$\sum_{i=1}^{K} p_i = 1 \quad (3)$$

Subsequently, the anticipated depth value of the pixel point can be calculated using the predicted probability distribution, which serves as the final depth estimation value. The formula for calculation is as follows:

$$D = \sum_{i=1}^{K} p_i \cdot d_i \quad (4)$$

$$d_i = \frac{b_i + b_{i+1}}{2} \quad (5)$$

The representative depth value of the i-th depth bin, denoted as d_i, is conventionally defined as the midpoint of the fixed bin, expressed mathematically as Equation 5.

The distinction in the depth interval attraction refinement strategy is found in the formulation of an adaptive attraction algorithm that progressively modifies the bin intervals to the left or right within the depth interval. Multi-scale features are employed to predict a set of points on the depth interval that will attract the bin center. The formula for the adjustment range of attraction is presented as follows:

$$\Delta c_i = \sum_{k=1}^{n_\alpha} \frac{a_k - c_i}{1 + \alpha \mid a_k - c_i \mid^{\gamma}} \quad (6)$$

In this context, C_i represents the original center point, while the hyperparameters α and γ dictate the strength of the attractor. Figure 3 illustrates the comprehensive structure of the Metric Bin Module.

B. *HFSPB*

A Heterogeneous Feature Single-Phase Booster (HFSPB) is proposed to augment high-dimensional feature extraction capacity, drawing inspiration from transformer architectures. This enhancement is particularly critical for monocular depth estimation tasks where intricate feature discrimination is paramount. Specifically, the HFSPB module amplifies feature discriminability in lightweight backbone networks through multi-scale receptive field expansion and cross-channel attention mechanisms, analogous to signal amplification processes in electrical systems.

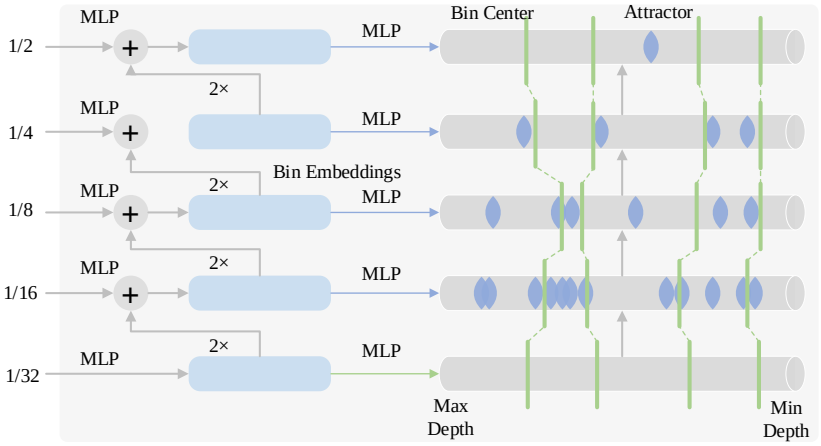

Fig. 3. Metric Bin Module

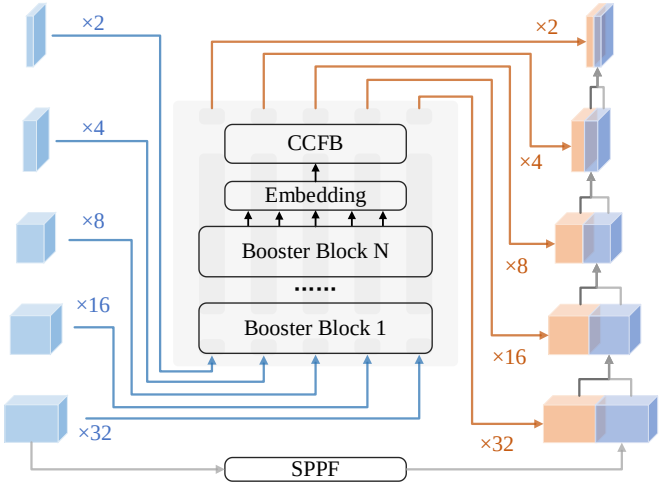

Fig. 4. Heterogeneous Feature Single-phase Booster

This module utilizes multi-scale feature information to enhance the network's feature extraction capabilities by incorporating the 2x, 4x, 8x, 16x, and 32x downsampled feature maps from the EIS-YOLO backbone, as well as the 32x maps following SPPF pyramid pooling. The Booster Block processes these through multiple rounds of feature extraction, with N established at 12 in our experiments. The Booster Block functions as a layer for feature extraction. Following an evaluation of multiple architectures, including ResNet[10], EfficientNet v2[11], FastViT[12], and EfficientFormer[13], RepViT[14] was selected for the Booster Block. The enhanced data subsequently passes through a Channel-wise Cross Fusion Block, modeled after UCTransNet[15], which integrates multi-scale data into transformer-format tokens and processes them using a channel-wise cross fusion transformer. This utilizes the capacity of transformers to capture long-range dependencies, thereby improving depth feature information across various scales. The output tokens undergo refinement via an attention module to enhance their feature representation. The multi-scale feature maps are concatenated with upsampled maps and input into the subsequent depth estimation head.

C. CCFB

The Channel-wise Cross Fusion Block, illustrated in Figure 4, enables the adaptive integration of multi-scale features from the Booster and the depth estimation head, thereby reducing noise and improving the retention of relevant information. This module consists of two components: the Channel-wise Cross Fusion Transformer (CCFT) and the Efficient Multiscale Attention (EMA). In contrast to UCTransNet, we utilize a singular CCFT operation. Our findings indicate that the presence of multiple Booster feature extraction layers in earlier stages renders additional CCFTs unnecessary, thereby increasing computational load without enhancing depth estimation accuracy. Additionally, we have implemented EMA[16], which, in contrast to the simpler CCFA attention of UCTransNet, mitigates the effects of dimensionality reduction during convolution and promotes a more uniform distribution of spatial semantic features, thus enhancing feature extraction capabilities.

The Multi-head Cross-Attention mechanism distinguishes itself from the traditional self-attention approach by functioning along the channel dimension instead of utilizing patches. Additionally, it incorporates instance normalization, which normalizes the similarity matrix for each instance within the similarity graph, thereby facilitating smooth gradient propagation. Consequently, with N attention heads, the output computation is performed as outlined in Equation 8. Following the application of a basic MLP and a residual operator, the output is derived according to Equation 9:

$$\begin{aligned} CA_i = M_i V^T &= \sigma\left[\psi\left(\frac{Q_i^T K}{\sqrt{C_\Sigma}}\right)\right] V^T \\ &= \sigma\left[\psi\left(\frac{W_{Q_i}^T T_i^T T_\Sigma W_K}{\sqrt{C_\Sigma}}\right)\right] W_V^T T_\Sigma^T \end{aligned} \tag{7}$$

$$MCA_i = \frac{CA_i^1 + CA_i^2 + , \ldots , + CA_i^N}{N} \tag{8}$$

$$O_i = \mathrm{MCA}_i + \mathrm{MLP}(\mathbf{Q}_i + \mathrm{MCA}_i) \tag{9}$$

The EMA reorganizes channels within the batch dimension, segmenting the channel space into multiple sub-features to enhance spatial semantic distribution and optimize feature extraction. In addition to encoding global information to optimize channel weights in parallel branches, EMA integrates the outputs of these branches through cross-dimensional interactions, thereby enhancing the precision of pixel-level relationships. The implementation of EMA attention enhances the extraction of essential pixel information for depth estimation tasks.

IV. Experimental Design and Implementation

A. Dataset Introduction

Due to the limited availability of datasets for estimating automotive damage depth, we trained our monocular depth estimation model using a large public dataset. The NYU Depth v2 dataset[17], developed by researchers at New York University with a Microsoft Kinect RGB and depth camera, includes 1449 pairs of densely annotated RGB images and their corresponding aligned depth images. The dataset encompasses over 1000 categories, featuring 464 new indoor scenes from three cities, 26 distinct scene types, and a total of 407,024 unmarked images, with each object categorized and assigned an instance number.

B. Experimental Setup

Experiments were conducted on an Ubuntu 20.04 cloud server, employing a Python 3.8 development environment alongside PyTorch 2.0.0 and CUDA 11.8. The computations were powered by the RTX 3090 GPU. We established 300 epochs for training, utilized a batch size of 16, and implemented Mosaic data augmentation, which was disabled during the final 10 training rounds. The input image dimensions were established at (640, 640), the optimizer utilized was AdamW, and the initial learning rate was set to 0.01.

In the training for depth estimation, all instance segmentation-specific layers following the Backbone were frozen, thereby training and immobilizing the complete instance segmentation task. This maintained stability in the instance segmentation task throughout the depth estimation training process.

C. Comparative Experiments

This research assessed the effects of different improvements in the monocular depth estimation component utilizing the NYU Depth v2 dataset. The branch integrated an HFSPB into the backbone to enhance feature extraction through an additional booster layer. The CCFB module in the booster enabled feature fusion and integration with the decoder, utilizing multi-scale data.

We evaluated various prominent lightweight backbone architectures utilizing convolutional neural networks and Vision Transformers for the Booster network selection. The findings are presented in Table I.

Among lightweight networks, Vision Transformers (ViT)

demonstrate superior performance compared to convolution-based models in depth estimation tasks, underscoring the necessity for effective feature extraction capabilities. Our experiments demonstrate that RepViT performs exceptionally in depth feature extraction, attaining δ metrics of 91.9%, 99.4%, and 99.8%, alongside the lowest REL and RMSE values.

In comparing our monocular depth estimation branch to prior models, as illustrated in Table II, our analysis reveals that although our ViT-based model exhibits a marginally slower inference time than the convolution-based BTS[18], it demonstrates a substantial improvement in depth estimation accuracy. In comparison to other ViT-based models such as AdaBins[19], LocalBins[20], and NeWCRFs[21], our method demonstrates both improved speed and marginally enhanced accuracy. In conclusion, while our model exhibits slightly lower accuracy than ZoeDepth, it offers an inference speed that is nearly three times faster, rendering it suitable for the lightweight and real-time requirements of automotive damage analysis.

D. Ablation Experiments

We conducted ablation experiments on the model to verify the effectiveness of our improvements, with results presented in Table III.

Directly appending a depth estimation head to the backbone resulted in suboptimal outcomes due to insufficient feature extraction, thereby undermining the accuracy of depth estimation. However, the addition of a Booster substantially enhanced feature extraction, leading to improved experimental outcomes and corroborating prior findings. The incorporation of the CCFB module improved the model's capacity to identify multi-scale key features, leading to a 5 percentage point increase in the δ metric. Transitioning from CCFA attention to EMA attention improved the outcomes, demonstrating the efficacy of the enhancements.

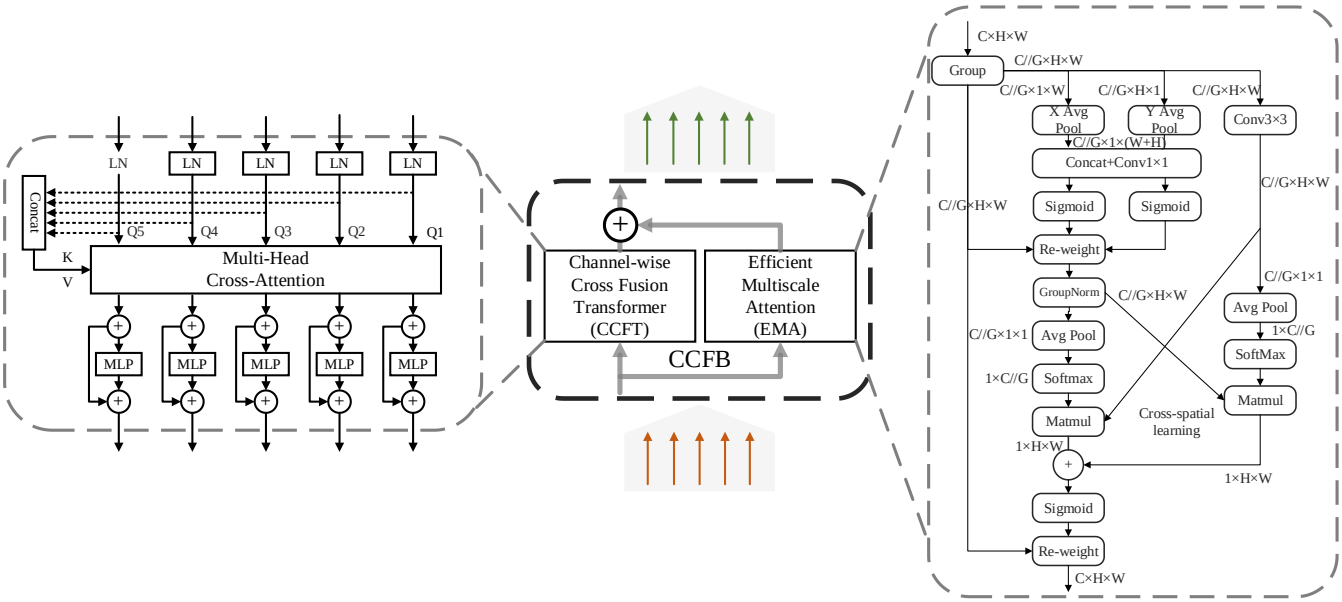

Fig. 5. Specification of CCFB

TABLE I. Experimental Results of Lightweight Feature Extraction Networks in Booster

Booster		δ_1	δ_2	δ_3	REL	RMSE	log10
CNN Based	+ResNet50	0.861	0.983	0.993	0.128	0.432	0.051
	+MobileNet v2	0.821	0.965	0.989	0.154	0.531	0.059
	+GhostNet	0.843	0.975	0.991	0.14	0.516	0.053
	+FasterNet	0.869	0.985	0.996	0.127	0.407	0.049
ViT Based	+MobileViT	0.863	0.983	0.995	0.129	0.412	0.05
	+FastViT	0.881	0.985	0.997	0.109	0.341	0.041
	+EfficientFormer	0.907	0.993	0.997	0.096	0.321	0.039
	+RepViT	0.919	0.994	0.998	0.089	0.312	0.037

TABLE II. Comparative Experiments with Other Models

Method	δ_1	δ_2	δ_3	REL	RMSE	log10	Inference time/s
BTS	0.883	0.977	0.994	0.112	0.392	0.048	6.2
AdaBins	0.901	0.983	0.997	0.105	0.369	0.044	20.3
LocalBins	0.903	0.986	0.998	0.102	0.358	0.042	20.5
NeWCRFs	0.911	0.991	0.998	0.098	0.334	0.041	29.8
ZoeD-X-N	0.946	0.995	0.999	0.082	0.294	0.035	33.2
ours	0.919	0.994	0.998	0.089	0.312	0.037	11.1

TABLE III. Ablation experiments

Booster	CCFT	CCFA	EMA	δ_1	δ_2	δ_3	REL	RMSE	log10
				0.722	0.942	0.981	0.191	0.693	0.067
√				0.851	0.981	0.993	0.133	0.439	0.053
√	√	√		0.903	0.994	0.998	0.107	0.326	0.041
√	√		√	0.919	0.994	0.998	0.089	0.312	0.037

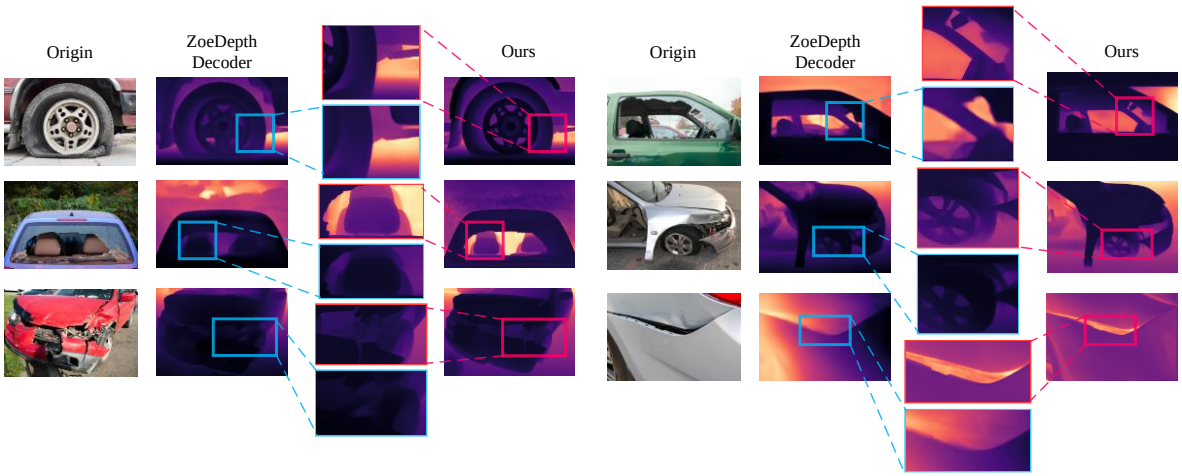

Fig. 6. Visual Contrast

E. Experimental Results Visualization

Our enhanced HFSPB structure, which incorporates a Booster layer with RepViT networks for improved feature extraction and a CCFB module for effective multi-scale feature integration, significantly enhances depth map visualization compared to the straightforward attachment of the ZoeDepth depth estimation head to the backbone, as demonstrated in the comparative Figure 6.

The refined depth head outputs exhibit enhanced depth information across various car details, demonstrating improved precision particularly at door frames, wheel hubs, and damage edges. This advancement facilitates more precise internal damage assessment, which is advantageous for subsequent damage analysis or efficient modeling processes.

F. Exploration of Vehicle Damage Analysis

Utilizing our multi-task model, we have accomplished instance segmentation for damage and obtained the corresponding depth data. This facilitates the integration of data with the previously mentioned sources, allowing for a systematic analysis of automotive accident damage. This is essential for intricate activities such as vehicle testing, insurance services, and accident scene reconstruction. Although a multi-task automotive damage dataset is absent, our algorithmic design enables sophisticated damage analysis.

Fig. 7. Examples of window damage

Traditional models for window damage primarily focus on segmenting the damaged areas to identify issues, which is inadequate for comprehensive analysis. With 3D information, it is possible to distinguish between a crack and penetration, evaluate the proportion of the damaged area relative to the component, and determine the shape of the damage.

Window damage typically impacts the front windshield, side windows, and rear windshield. Damage generally advances from cracks to shattering and subsequently to breaking. Instance segmentation is capable of detecting damaged windows; however, it is unable to perform assessments regarding their condition. Depth maps convey this information via variations in depth intervals, as illustrated in Figure 7. This highlights the significance of 3D data in the analysis of automotive damage.

Initially, we input images of the damaged car into a multi-task model for specific operations. This model provides precise instance segmentation of the car's damaged areas, producing contours and positional data for each affected region. Simultaneously, it produces depth maps that represent the depth information of each pixel, facilitating the comprehension of the three-dimensional structure of the damage. Using instance segmentation results, we isolate the window glass components from the depth map, a critical step that focuses the analysis on the specific window glass area while excluding irrelevant regions. Subsequently, we utilize image processing algorithms to analyze the depth map, focusing on pixel depth information to detect anomalies within the window. By calculating the maximum and minimum depths of the window area and comparing the difference, one can infer the presence of a hole if the difference exceeds a predetermined threshold. Based on this inference, subsequent image processing utilizing a region growing algorithm delineates potential hole areas on the depth map. This method determines the presence of a hole and accurately delineates its shape and size, enabling the calculation of the damage proportion. Figure 8 illustrates the entire process. We successfully calculate the window hole rate and determine the shape of the hole.

V. Conclusion

This study utilizes the YOLOv8 benchmark and extends previous research to create a monocular depth estimation branch for analyzing automotive damage. The depth estimation head employs a dynamic strategy for adjusting depth ranges, thereby enhancing the accuracy of depth predictions. We propose the incorporation of an HFSPB feature extraction module between the depth head and the backbone network to facilitate the integration of instance segmentation with depth estimation. This module incorporates a Booster layer utilizing RepViT and a CCFB module with CCFT and EMA, thereby improving multi-scale feature extraction and enhancing the accuracy of depth prediction. The model attains δ metrics of 91.9%, 99.4%, and 99.8%, demonstrating minimal relative error (REL) and root mean square error (RMSE). Our model provides faster inference, reduced computational costs, and improved edge deployment capabilities compared to existing models, effectively addressing the requirements of automotive damage analysis.

Our multi-task approach to automotive damage analysis has revealed deeper insights and methodologies for specific damages, providing a foundation for future end-to-end solutions in this area.

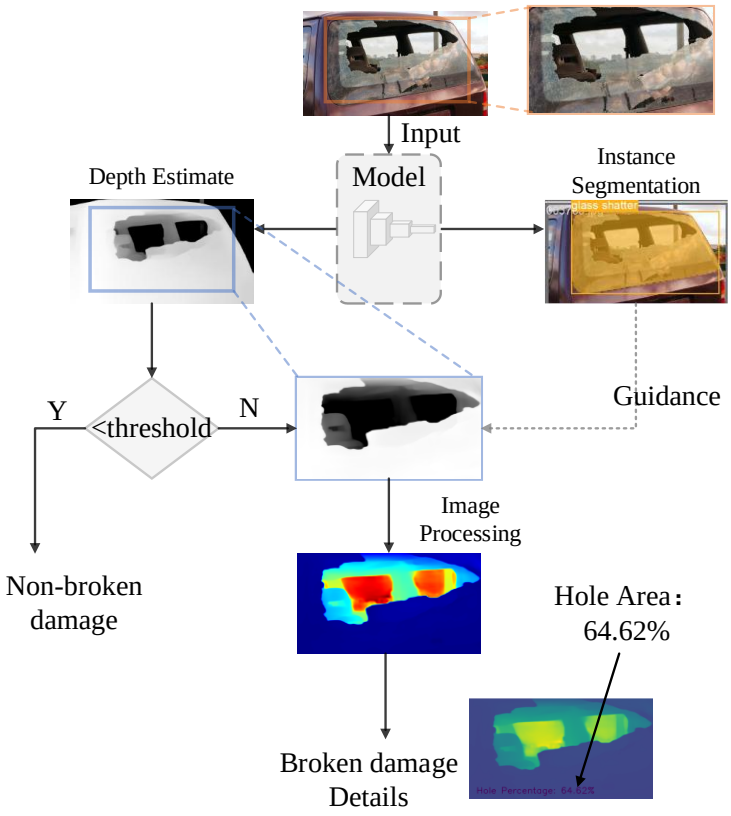

Fig. 8. Comprehensive analytical flowchart for vehicular window impairment

REFERENCES

[1] C. J. Jiao. "Research on image defect detection of tire X-Ray based on Faster R-CNN," Shan Dong: Qing Dao University of Technology, 2020.

[2] C. P. Zhu and Y. B. Yang. "Online detection algorithm of automobile wheel surface defects based on improved Faster-RCNN model," in Surface Technology, vol. 49, no. 6, pp. 359-365, 2020.

[3] F. R. Sun, N. Xiao and Y. X. Wu. "Defect segmentation algorithm of auto parts based on non-local U-Net model," in Electronic Design Engineering, vol. 30, no. 16, pp. 70-74, 2022.

[4] C. W. Tang. "Semantic segmentation network for surface defect detection of automobile wheel hub fusing high-resolution feature and multi-scale feature," in Applied Sciences, vol. 11, no. 22, p. 10508.

[5] Z. Y. Zhang, Y. Liu and F. C. Liu. "Research and analysis of automobile wheel surface defect detection algorithm based on YOLOv3-spp," in Acta Metrologica Sinica, vol. 44, 2023, pp. 1375-1382.

[6] H. Lv. "Automotive glass defect detection based on yolo algorithm," Fu Jian: Fujian University of Technology, 2023.

[7] X. Tan and J. Zhao. "Enhancing YOLOv8 for Improved Instance Segmentation of Automotive Surface Damage," in Journal of Computer Engineering & Applications, vol. 60, no. 14, 2024.

[8] X. K. Wang, W. J. Li and Z. C. Wu. "Cardd: A new dataset for vision-based car damage detection," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 7, pp. 7202-7214, 2023.

[9] Bhat and S. Farooq. "Zoedepth: Zero-shot transfer by combining relative and metric depth," arXiv preprint arXiv:2302.12288, 2023.

[10] K. He and X. Zhang. "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[11] M. Tan. "Efficientnet: Rethinking model scaling for convolutional neural networks," arXiv preprint arXiv:1905.11946, 2019.

[12] P. K. A. Vasu, J. Gabriel, and J. Zhu. "FastViT: A fast hybrid vision transformer using structural reparameterization," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5785-5795.

[13] Y. Li, G. Yuan, and Y. Wen. "Efficientformer: Vision transformers at mobilenet speed (Proceedings style)," in Advances in Neural Information Processing Systems, vol. 35, pp. 12934-12949, 2022.

[14] A. Wang, H. Chen, and Z. Lin. "Repvit: Revisiting mobile cnn from vit perspective," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15909-15920.

[15] H. Wang, P. Cao, J. Wang. "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 3, pp. 2441-2449, 2022.

[16] D. Ouyang, S. He, and G. Zhang. "Efficient multiscale attention module with cross-spatial learning," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5.

[17] N. Silberman, D. Hoiem, and P. Kohli. "Indoor segmentation and support inference from rgbd images," in Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer Berlin Heidelberg, 2012, pp. 746-760.

[18] J. H. Lee, M. K. Han, and D. W. Ko. "From big to small: Multi-scale local planar guidance for monocular depth estimation," arXiv preprint arXiv:1907.10326, 2019.

[19] S. F. Bhat, I. Alhashim, and P. Wonka. "Adabins: Depth estimation using adaptive bins," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4009-4018.

[20] S. F. Bhat, I. Alhashim, and P. Wonka. "Localbins: Improving depth estimation by learning local distributions," in European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022, pp. 480-496.

[21] W. Yuan, X. Gu, and Z. Dai. "Newcrfs: Neural window fully-connected crfs for monocular depth estimation. arXiv 2022," arXiv preprint arXiv:2203.01502, 2022.