CAR DAMAGE DETECTION AND CAPTION GENERATION USING DEEP LEARNING

Priyadarshini Munigala, Sri Sai Lalitha Mallika Yeturi
Yeshiva University
pmunigal@mail.yu.edu, syeturi@mail.yu.edu

Abstract

Car damage detection and caption generation play a vital role in expediting insurance claims, vehicle inspections, and automotive repair processes. This paper introduces an end-to-end deep learning-based framework that integrates a YOLOv7 object detection model with a Transformer-based natural language processing (NLP) module to localize and describe car damages automatically. We present our motivation for automating vehicle inspections, review the related work in both computer vision and NLP, and detail our dataset preparation, model architecture, training pipeline, and experimental results. Our system demonstrates high accuracy in detecting various damage categories, such as dents, scratches, and crushed areas, alongside generating coherent and contextually relevant captions. The proposed approach offers a scalable and reliable solution, minimizing manual labor and improving consistency in damage assessment across the automotive industry.

Keywords: Car Damage Detection, YOLO, Transformers, Deep Learning, NLP, Image Captioning

1. Introduction

The automotive industry has evolved rapidly over the last decade, with unprecedented growth in the number of vehicles on the road. This surge has amplified the need for efficient and accurate vehicle inspection methods, particularly in areas such as insurance claims, fleet management, and customer returns in car rental services. Traditional vehicle inspections often rely on manual assessments conducted by human inspectors, who may be limited by time constraints and prone to subjective judgment. As a result, insurers and repair companies face challenges in handling large volumes of claims, which can lead to backlogs, slower processing times, and inconsistencies in damage evaluation.

Recent advances in artificial intelligence (AI) and computer vision have showcased the feasibility of automating labor-intensive tasks. Deep learning models, in particular, have proven successful in complex visual recognition tasks, providing robust performance in object detection, classification, and segmentation. Early methods like Fast/Faster R-CNN [5] and YOLO [4] effectively identify objects within an image but do not inherently provide explanatory or descriptive output. This limitation poses a challenge in real-world scenarios, where insurers or mechanics benefit not only from localized damages but also from text-based, context-rich descriptions.

To address this gap, researchers have explored image captioning, where models generate textual descriptions from visual inputs. Transformers [6] have significantly pushed the state-of-the-art by capturing long-range dependencies and dynamic context. However, these solutions typically focus on general images and do not incorporate domain-specific vocabularies—critical for tasks such as car damage assessment.

In this paper, we propose a unified system that integrates YOLOv7 [8] for car damage detection with a Transformer-based NLP module for caption generation. By providing both bounding box annotations of damages and corresponding textual descriptions, our pipeline streamlines insurance claim handling, repair diagnostics, and even large-scale fleet management. We further detail a comprehensive dataset curated from real-world insurance repositories and user-submitted photos, emphasizing diverse damage types and environmental conditions. Our experiments show that this end-to-end framework not only automates detection but also yields coherent and informative captions, laying a strong foundation for next-generation vehicle inspection technologies.

1.1. Motivation

Manual vehicle inspection can be:

- **Time-Intensive:** Repetitive tasks require significant labor, delaying claim settlements and repair estimates.
- **Subjective:** Evaluations depend heavily on inspector experience, leading to varied outcomes for similar

damages.

- **Costly:** High volumes of inspections strain resources, particularly in large-scale fleet operations or massive insurance pools.

By introducing an automated AI-driven approach, these bottlenecks can be alleviated, enabling standardized assessments, consistent textual reports, and reduced overhead costs.

2. Related Work

2.1. Vehicle Damage Detection

Computer vision has seen rapid growth in object detection capabilities. Lin *et al.* [3] introduced focal loss for addressing class imbalance, enhancing detection performance for underrepresented categories. YOLO-based architectures [4, 8] further refined real-time detection, focusing on more efficient backbones and optimized anchors. Deformable convolutions have also shown promise in dealing with irregularly shaped damage regions, as suggested by Zhu *et al.* [10].

2.2. Caption Generation

Caption generation has undergone a paradigm shift from RNN-based encoder-decoder models [9] to Transformer architectures [6]. The self-attention mechanism allows for parallel computation and fine-grained contextual relationships, leading to state-of-the-art results in various language tasks. However, captioning models tailored to specialized domains—such as automobile inspection—are less common, leaving an opportunity for significant contributions in bridging structured vision outputs and natural language descriptions.

2.3. Combined Detection-Captioning Frameworks

Faster R-CNN [5] and Show-and-Tell [7] exemplify separate yet powerful approaches for object detection and caption generation, respectively. Existing integrated frameworks often address general images without focusing on domain-specific requirements. The lack of specialized vocabularies or context in damage descriptions underscores a research gap that this paper aims to fill by merging YOLOv7 detection outputs with a Transformer-based descriptive module. Methods like Fast R-CNN [1] could also be extended to incorporate textual generation capabilities, particularly for region-based damage assessments.

3. Methodology

3.1. Dataset Description

Our dataset consists of 10,000 high-resolution images of vehicles, curated from various sources, including insurance repositories, user-submitted photos, and controlled lab environments. Each image is annotated with bounding boxes for different damage types (*dents, scratches, crushed parts*) and paired with a short textual description.

3.1.1 Preprocessing Steps

- **Resizing to** 416×416 pixels to align with YOLOv7's input size.
- **Normalization** of pixel intensities to [0,1] for numerical stability.
- **Augmentation**: Horizontal flips (50%), rotations up to $\pm20^\circ$, and brightness variations up to $\pm20\%$ to simulate real-world conditions.

These steps mitigate overfitting, enhance domain variance, and improve the model's ability to generalize to unseen scenarios.

3.2. Model Architecture

YOLOv7 [8] is the primary detection module, responsible for identifying and localizing damages in real time. For each detected bounding box, we extract regional features that feed into our Transformer-based caption generator [6]. This design ensures that the language model has direct access to domain-relevant features corresponding to specific damages. Model compression strategies, such as AMC (AutoML for Model Compression) [2], can further optimize real-time performance on edge devices.

3.3. Training Pipeline

We train the YOLOv7 detector using a multi-component loss function consisting of:

- **Classification Loss** for damage category prediction (dent, scratch, crushed).
- **Localization Loss** (often a form of IoU or GIoU) for precise bounding box regression.
- **Objectness Loss** for robust confidence estimation.

The Transformer model is trained on cross-entropy loss between predicted captions and ground-truth descriptions. Both modules utilize the Adam optimizer with an initial learning rate of 10^{-4} and mini-batch size set to 16. Training spans 50 epochs, with early stopping based on validation metrics (mAP for detection and BLEU-4 for captioning).

4. Results and Discussion

4.1. Object Detection Performance

Table 3 details the class-wise results obtained by our YOLOv7-based detection module. Overall, the mean av-

erage precision (mAP) reached 87%, outperforming alternative frameworks like Faster R-CNN [5] (82% mAP) and RetinaNet [3] (84% mAP). The model efficiently handles various angles and lighting conditions, though highly reflective surfaces can still pose difficulties.

Table 1. Object Detection Results

Category	Precision	Recall
Dents	89%	85%
Scratches	86%	83%
Crushed	88%	87%

4.2. Caption Generation Performance

For the caption generation component, we employ BLEU-4, METEOR, and ROUGE-L metrics to evaluate linguistic quality. Table 4 shows our results, indicating contextually accurate and coherent outputs:

- **BLEU-4 of 0.81**: Demonstrates close alignment with reference captions.
- **METEOR of 0.68**: Signifies robust matching at the synonym and phrase level.
- **ROUGE-L of 0.74**: Highlights the coverage of essential n-grams from ground truths.

Qualitative evaluations further confirm the model's ability to describe damage severity and location, which can enhance communication between insurance adjusters, repair technicians, and vehicle owners.

Table 2. Caption Generation Metrics

Metric	Score
BLEU-4	0.81
METEOR	0.68
ROUGE-L	0.74

4.3. Error Analysis

Errors typically emerge in two primary scenarios:

1. **Occlusions & Complex Backgrounds**: Overlapping objects or cluttered backgrounds can lead to misclassification, particularly for subtle damages such as light scratches.
2. **Caption Specificity**: Some generated captions can be overly general or fail to capture minor yet relevant details (e.g., small paint chips), suggesting the need for more fine-grained features and possibly domain-specific vocabulary expansions.

Future efforts may involve training the Transformer with multi-scale features extracted at different resolution levels, as well as augmenting the dataset with more diverse contexts.

5. Comparative Analysis

5.1. Alternative Detection Approaches

Faster R-CNN [5] and RetinaNet [3] remain competitive, especially for smaller datasets. However, YOLOv7 demonstrates a crucial balance of speed and accuracy, making it better suited for real-time applications in large-scale insurance workflows or on-device deployments, where rapid processing is paramount. Deformable DETR [10] has also shown potential for handling shape variance, which could be explored for car damage scenarios.

5.2. Caption Generation Strategies

While LSTM/GRU-based encoders [9] still see use in some production systems, the Transformer's ability to leverage self-attention for contextual modeling has proven superior in generating coherent domain-specific captions. This advantage is especially pronounced in automotive damage reports that demand detailed descriptions rather than broad, generic statements.

5.3. Integration Challenges

Bridging object detection and captioning into a single pipeline necessitates careful synchronization between bounding box predictions and linguistic tokens. Latency can be introduced if the detected features are not efficiently transferred to the Transformer model. Techniques such as knowledge distillation or pruning [2] could further optimize inference speed, particularly for edge-based scenarios.

6. Ethical Considerations

6.1. Bias in AI Systems

AI models are susceptible to biases if the underlying dataset underrepresents certain vehicle types, damage categories, or environmental conditions. To ensure fairness, data collection strategies must be periodically audited, incorporating diverse real-world images that span various vehicle models, colors, and geographic regions.

6.2. Privacy Implications

Captured vehicle images can reveal personal data (e.g., license plates, interior items). Compliance with data protection regulations such as the General Data Protection Regulation (GDPR) mandates anonymization and clear user consent. Potential methods include masking sensitive regions or employing in-vehicle edge processing to minimize data sharing.

6.3. Regulatory Compliance

With AI models increasingly guiding insurance decisions or safety assessments, regulatory oversight must ensure transparency and accountability. Explainable AI techniques, such as visualizing attention maps, can bolster stakeholder trust by elucidating how damage detection and caption generation decisions are reached.

7. Future Work

7.1. Edge-Case Handling and Domain Expansion

While our dataset covers common damages, rare or extreme scenarios (e.g., vehicles in floods, fires, or severe collisions) remain underexplored. Future research might include expanding the data to capture a broader spectrum of damage severity and leveraging semi-supervised or unsupervised methods for continuous learning in evolving real-world conditions.

7.2. Online and Edge Deployment

To ensure scalability, deploying the model on mobile or embedded devices (e.g., in-car systems) can facilitate real-time damage evaluations without relying on cloud infrastructures. Model compression, pruning, and quantization techniques would be crucial to meet the computational limitations of edge devices while maintaining accuracy.

7.3. Multimodal Inputs and Multilingual Captioning

Integrating additional sensors, such as LiDAR or thermal imaging, could provide richer contextual information, notably for damages obscured by low visibility or poor lighting. Moreover, extending caption generation to multiple languages would enhance accessibility in multinational fleets or insurance companies, requiring cross-lingual or multilingual Transformer training.

7.4. Explainability and User Interaction

Promoting user trust calls for interpretable results. Visualizing YOLOv7 activation maps and Transformer attention layers can clarify why certain regions trigger detection or specific descriptive phrases. Interactive interfaces might allow users to refine automated captions or correct bounding box predictions, enabling a human-AI collaborative workflow.

8. Related Work

8.1. Vehicle Damage Detection

Computer vision has seen rapid growth in object detection capabilities. Lin *et al.* [3] introduced focal loss for addressing class imbalance, enhancing detection performance for underrepresented categories. YOLO-based architectures [4, 8] further refined real-time detection, focusing on more efficient backbones and optimized anchors. Such improvements provide the bedrock for accurate, domain-specific tasks like automotive damage detection.

8.2. Caption Generation

Caption generation has undergone a paradigm shift from RNN-based encoder-decoder models [9] to Transformer architectures [6]. The self-attention mechanism allows for parallel computation and fine-grained contextual relationships, leading to state-of-the-art results in various language tasks. However, captioning models tailored to specialized domains—such as automobile inspection—are less common, leaving an opportunity for significant contributions in bridging structured vision outputs and natural language descriptions.

8.3. Combined Detection-Captioning Frameworks

Faster R-CNN [5] and Show-and-Tell [7] exemplify separate yet powerful approaches for object detection and caption generation, respectively. Existing integrated frameworks often address general images without focusing on domain-specific requirements. The lack of specialized vocabularies or context in damage descriptions underscores a research gap that this paper aims to fill by merging YOLOv7 detection outputs with a Transformer-based descriptive module.

9. Methodology

9.1. Dataset Description

Our dataset consists of 10,000 high-resolution images of vehicles, curated from various sources, including insurance repositories, user-submitted photos, and controlled lab environments. Each image is annotated with bounding boxes for different damage types (*dents, scratches, crushed parts*) and paired with a short textual description.

9.1.1 Preprocessing Steps

- **Resizing to** 416×416 pixels to align with YOLOv7's input size.
- **Normalization** of pixel intensities to [0,1] for numerical stability.
- **Augmentation**: Horizontal flips (50%), rotations up to $\pm20^\circ$, and brightness variations up to $\pm20\%$ to simulate real-world conditions.

These steps mitigate overfitting, enhance domain variance, and improve the model's ability to generalize to unseen scenarios.

9.2. Model Architecture

YOLOv7 [8] is the primary detection module, responsible for identifying and localizing damages in real time. For each detected bounding box, we extract regional features that feed into our Transformer-based caption generator [6]. This design ensures that the language model has direct access to domain-relevant features corresponding to specific damages.

9.3. Training Pipeline

We train the YOLOv7 detector using a multi-component loss function consisting of:

- **Classification Loss** for damage category prediction (dent, scratch, crushed).
- **Localization Loss** (often a form of IoU or GIoU) for precise bounding box regression.
- **Objectness Loss** for robust confidence estimation.

The Transformer model is trained on cross-entropy loss between predicted captions and ground-truth descriptions. Both modules utilize the Adam optimizer with an initial learning rate of 10^{-4} and mini-batch size set to 16. Training spans 50 epochs, with early stopping based on validation metrics (mAP for detection and BLEU-4 for captioning).

10. Results and Discussion

10.1. Object Detection Performance

Table 3 details the class-wise results obtained by our YOLOv7-based detection module. Overall, the mean average precision (mAP) reached 87%, outperforming alternative frameworks like Faster R-CNN [5] (82% mAP) and RetinaNet [3] (84% mAP). The model efficiently handles various angles and lighting conditions, though highly reflective surfaces can still pose difficulties.

Table 3. Object Detection Results

Category	Precision	Recall
Dents	89%	85%
Scratches	86%	83%
Crushed	88%	87%

10.2. Caption Generation Performance

For the caption generation component, we employ BLEU-4, METEOR, and ROUGE-L metrics to evaluate linguistic quality. Table 4 shows our results, indicating contextually accurate and coherent outputs:

- **BLEU-4 of 0.81**: Demonstrates close alignment with reference captions.
- **METEOR of 0.68**: Signifies robust matching at the synonym and phrase level.
- **ROUGE-L of 0.74**: Highlights the coverage of essential n-grams from ground truths.

Qualitative evaluations further confirm the model's ability to describe damage severity and location, which can enhance communication between insurance adjusters, repair technicians, and vehicle owners.

Table 4. Caption Generation Metrics

Metric	Score
BLEU-4	0.81
METEOR	0.68
ROUGE-L	0.74

10.3. Error Analysis

Errors typically emerge in two primary scenarios:

1. **Occlusions & Complex Backgrounds**: Overlapping objects or cluttered backgrounds can lead to misclassification, particularly for subtle damages such as light scratches.
2. **Caption Specificity**: Some generated captions can be overly general or fail to capture minor yet relevant details (e.g., small paint chips), suggesting the need for more fine-grained features and possibly domain-specific vocabulary expansions.

Future efforts may involve training the Transformer with multi-scale features extracted at different resolution levels, as well as augmenting the dataset with more diverse contexts.

11. Comparative Analysis

11.1. Alternative Detection Approaches

Faster R-CNN [5] and RetinaNet [3] remain competitive, especially for smaller datasets. However, YOLOv7 demonstrates a crucial balance of speed and accuracy, making it better suited for real-time applications in large-scale insurance workflows or on-device deployments, where rapid processing is paramount.

11.2. Caption Generation Strategies

While LSTM/GRU-based encoders [9] still see use in some production systems, the Transformer's ability to leverage self-attention for contextual modeling has proven superior in generating coherent domain-specific captions. This advantage is especially pronounced in automotive damage reports that demand detailed descriptions rather than broad, generic statements.

11.3. Integration Challenges

Bridging object detection and captioning into a single pipeline necessitates careful synchronization between bounding box predictions and linguistic tokens. Latency can be introduced if the detected features are not efficiently transferred to the Transformer model. Techniques such as knowledge distillation or pruning could further optimize inference speed, particularly for edge-based scenarios.

12. Ethical Considerations

12.1. Bias in AI Systems

AI models are susceptible to biases if the underlying dataset underrepresents certain vehicle types, damage categories, or environmental conditions. To ensure fairness, data collection strategies must be periodically audited, incorporating diverse real-world images that span various vehicle models, colors, and geographic regions.

12.2. Privacy Implications

Captured vehicle images can reveal personal data (e.g., license plates, interior items). Compliance with data protection regulations such as the General Data Protection Regulation (GDPR) mandates anonymization and clear user consent. Potential methods include masking sensitive regions or employing in-vehicle edge processing to minimize data sharing.

12.3. Regulatory Compliance

With AI models increasingly guiding insurance decisions or safety assessments, regulatory oversight must ensure transparency and accountability. Explainable AI techniques, such as visualizing attention maps, can bolster stakeholder trust by elucidating how damage detection and caption generation decisions are reached.

13. Future Work

13.1. Edge-Case Handling and Domain Expansion

While our dataset covers common damages, rare or extreme scenarios (e.g., vehicles in floods, fires, or severe collisions) remain underexplored. Future research might include expanding the data to capture a broader spectrum of damage severity and leveraging semi-supervised or unsupervised methods for continuous learning in evolving real-world conditions.

13.2. Online and Edge Deployment

To ensure scalability, deploying the model on mobile or embedded devices (e.g., in-car systems) can facilitate real-time damage evaluations without relying on cloud infrastructures. Model compression, pruning, and quantization techniques would be crucial to meet the computational limitations of edge devices while maintaining accuracy.

13.3. Multimodal Inputs and Multilingual Captioning

Integrating additional sensors, such as LiDAR or thermal imaging, could provide richer contextual information, notably for damages obscured by low visibility or poor lighting. Moreover, extending caption generation to multiple languages would enhance accessibility in multinational fleets or insurance companies, requiring cross-lingual or multilingual Transformer training.

13.4. Explainability and User Interaction

Promoting user trust calls for interpretable results. Visualizing YOLOv7 activation maps and Transformer attention layers can clarify why certain regions trigger detection or specific descriptive phrases. Interactive interfaces might allow users to refine automated captions or correct bounding box predictions, enabling a human-AI collaborative workflow.

14. Conclusion

This study presents a comprehensive end-to-end system for car damage detection and descriptive caption generation, combining YOLOv7's robust real-time object detection with a Transformer-based NLP module. Our approach significantly streamlines vehicle inspection processes by providing both bounding box annotations and textual descriptions, thereby expediting insurance claim handling, repair shop diagnostics, and large-scale fleet management. Experimental evaluations indicate that the system achieves an mAP of 87% for object detection and strong BLEU, METEOR, and ROUGE scores for caption generation, underscoring its real-world viability.

Moreover, the synergy between robust object detection and domain-specific captioning opens doors for advanced analytics, such as automated severity assessment and part-specific analysis. By capturing the nuanced language of damage descriptions, the Transformer module can be fine-tuned to detect emerging vehicle technologies, custom modifications, or novel damage types. In addition, integrating explainable AI (XAI) techniques within the workflow can further build confidence among stakeholders, including insurance adjusters, mechanics, and end-users.

Looking ahead, future work involves addressing tricky edge cases (e.g., extremely high damage severity, unusual environmental conditions) and enhancing on-device performance through hardware-aware optimizations. Continuous model adaptation, driven by semi-supervised techniques and data streaming, could also ensure that the framework remains robust as automotive designs evolve. By combining these avenues of research, our proposed system can re-

main at the forefront of automated car damage assessment, ultimately reducing operational costs, improving safety, and enhancing customer satisfaction.

References

[1] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[2] Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. AMC: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018. 2, 3

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2, 3, 4, 5

[4] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2, 4

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015. 1, 2, 3, 4, 5

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 1, 2, 4, 5

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. 2, 4

[8] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 1, 2, 4, 5

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015. 2, 3, 4, 5

[10] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3