DATA645 Neural Network and Deep Learning

Final Project Report

Group Member: Ziliang Song, Dongni Li, Pin Tzu Tseng

Abstract

Semantic segmentation enables pixel-level understanding in autonomous driving systems. This project conducts a quantitative comparison of three U-Net training strategies, baseline (random initialization), data augmentation, and transfer learning to evaluate their effects on accuracy, robustness, and training efficiency. Experiments on the Carvana dataset reveal that the Baseline model achieves a high Dice score of 0.9916 yet suffers from strong validation-loss spikes. Data augmentation stabilizes training but slightly reduces IoU (0.9833 → 0.9746). Transfer learning using a VGG16 encoder achieves the best overall performance, reaching 0.9936 Dice, 0.9872 IoU, and a 26% speed improvement. These results confirm that U-Net with transfer learning is the most effective strategy for high-precision vehicle segmentation tasks.

1. Introduction

Semantic segmentation is a fundamental computer vision task that assigns a semantic label to every pixel in an image. In autonomous driving, precise pixel-level understanding is required for vehicle detection, road boundary identification, obstacle recognition, and scene parsing. U-Net, with its symmetric encoder–decoder structure and skip connections, has become one of the most prominent architectures for segmentation tasks requiring fine boundary reconstruction.

Despite U-Net's strong baseline performance, its training behavior and final accuracy depend heavily on the chosen training strategy. Random initialization may cause unstable gradients and slow convergence. Data augmentation can improve generalization but may distort fine-grained details. Transfer learning can provide more stable feature extraction, yet it changes the architecture and may affect computational efficiency.

This project investigates the following research questions:

- **Robustness:** Can a randomly initialized U-Net (Baseline) maintain stable convergence during training?
- **Trade-off:** Can data augmentation enhance robustness without significantly reducing segmentation accuracy?
- **Optimal strategy:** Can transfer learning simultaneously improve accuracy, robustness, and training efficiency?

To answer these questions, we systematically evaluate three U-Net training strategies on the Carvana vehicle-segmentation dataset. Our goal is to provide empirical evidence identifying the best overall approach for high-precision segmentation tasks.

2. Related Work

2.1 Semantic Segmentation Architectures

Semantic segmentation is a pivotal technique in computer vision. SegNet, proposed by Badrinarayanan et al. (2017), introduced a deep convolutional encoder-decoder architecture that is particularly memory-efficient. A key feature of SegNet is its use of max-pooling indices transferred from the encoder to the decoder to perform non-linear upsampling, which eliminates the need for learning upsampling parameters while retaining boundary details (Badrinarayanan et al., 2017). While SegNet set a strong foundation, U-Net and its variants have become the dominant choice for tasks requiring fine-grained localization. Arulananth et al. (2024) demonstrated the effectiveness of the U-Net model specifically for urban environment analysis, showing its capability to handle complex imagery with high precision.

2.2 Training Strategies and Model Enhancements

Beyond architectural innovations, training strategies play a crucial role in model performance. Dimitrovski et al. (2024) explored the use of U-Net Ensembles for semantic segmentation in remote sensing. Their work suggests that combining predictions from multiple models, initialized differently or trained on different data subsets can significantly improve segmentation accuracy and generalization compared to a baseline model (Dimitrovski et al., 2024). This finding aligns with the broader research objective of identifying optimal training protocols, such as the comparison between random initialization, data augmentation, and transfer learning investigated in this study. Transfer Learning leverages pretrained encoders to provide rich initialization. As noted in comprehensive surveys on autonomous driving perception, utilizing pretrained models is a standard practice to accelerate convergence and improve robustness, especially when domain-specific data is limited (Feng et al., 2020).

3. Methodology

3.1 Dataset and Evaluation Metrics

The Carvana Image Masking Challenge dataset consists of 5,086 high-resolution images of vehicles. We split the dataset into:

- Training set: 3,560
- Validation set: 763
- Test set: 763

Evaluation metrics include:

- Intersection-over-Union (IoU)
- Dice Coefficient (F1 Score)
- Pixel Accuracy
- Binary Cross-Entropy Loss

3.2 Model Architectures

Baseline U-Net

A standard four-level encoder–decoder architecture initialized with random weights.

Optimizer: Adam (learning rate = 1e-4)

Regularization: Early stopping with patience = 5

Augmented U-Net

The same architecture as the baseline, but trained with the following image augmentations to improve robustness:

Random horizontal and vertical flips

Random rotations with a rotation factor of 0.2

Photometric adjustments such as brightness and contrast variations

Transfer Learning U-Net (VGG16 Encoder)

A U-Net variant that uses a VGG16 backbone pretrained on ImageNet as the encoder:

Frozen layers: The first 15 VGG16 layers are kept non-trainable to preserve generic low-level features.

Trainable layers: Deeper convolutional layers and the bottleneck are fine-tuned for the Carvana task.

Decoder: The decoder is randomly initialized and trained end-to-end.

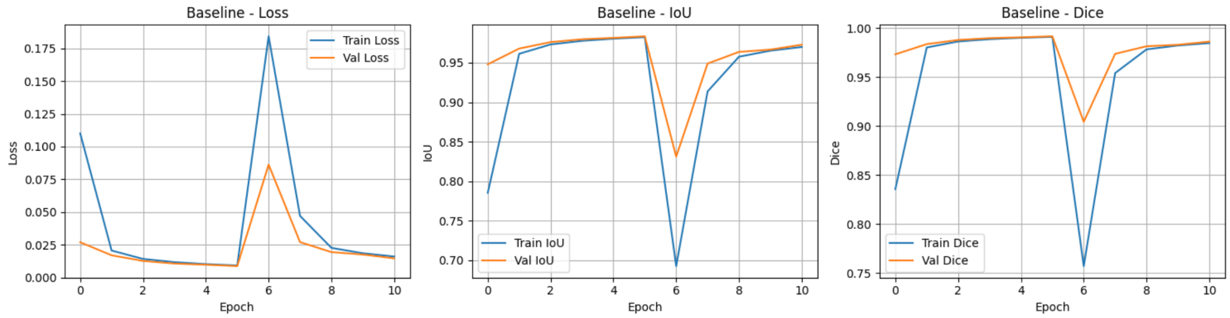
Skip connections: Feature maps from block1_conv2, block2_conv2, block3_conv3, and block4_conv3 are connected to the corresponding decoder stages.

4. Results

4.1 Baseline Model Performance

The Baseline U-Net demonstrated strong quantitative performance, achieving a Dice score of 0.9916 and an IoU of 0.9833. However, the training process revealed signs of instability. As shown in Figure 1, the validation loss exhibits a dramatic spike around epoch 7 and both IoU and Dice drop sharply at the same time. This anomaly suggests potential issues with gradient stability or learning-rate sensitivity during training, despite the model eventually converging to a high-performance state.

Figure 1. Baseline model training and validation curves for loss, IoU, and Dice.

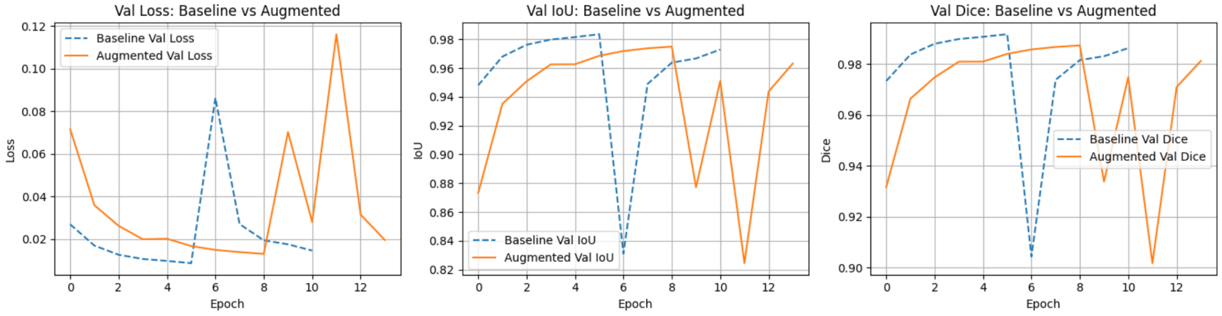


4.2 Augmented Model Performance

While the Augmented U-Net was expected to enhance robustness, the results reveal a clear trade-off. As illustrated in Figure 2, data augmentation reduces the severe gradient instability observed in the early Baseline epochs, but it does not lead to a consistently smooth convergence. The validation loss curve shows a noticeable spike around epoch 11, suggesting that the model struggled with certain augmented samples near convergence.

In terms of quantitative performance, the Augmented model achieves an IoU of 0.9746 and a Dice score of 0.9871. This represents a slight degradation compared to the Baseline model (IoU 0.9833), indicating that although augmentation introduces useful diversity, it may also add noise that hinders fine-grained boundary recovery.

Figure 2. Training and Validation loss, IoU, and Dice comparison between the Baseline and Augmented U-Net models.

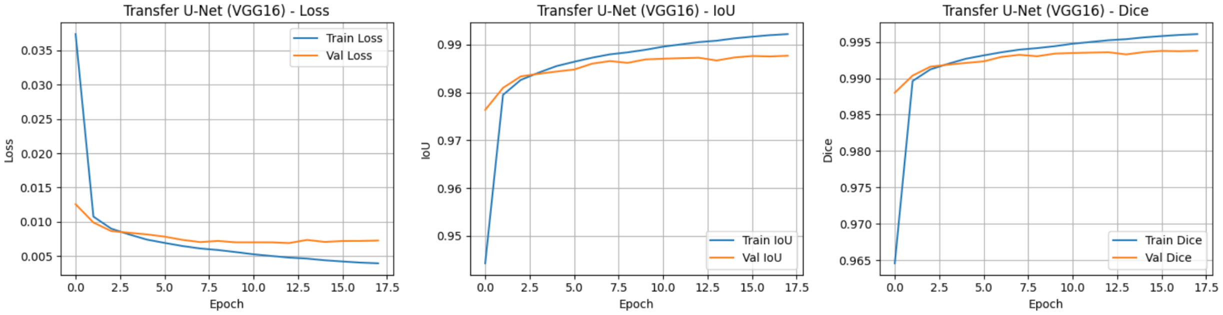


4.3 Transfer Learning Model Performance

The VGG16-based U-Net achieves the best overall performance among the three training strategies. On the test set, it reaches a Dice score of 0.9936 and an IoU of 0.9872, which are the highest values across all models. In addition, the average training time per step is reduced by about 26% compared to the Baseline model, indicating that transfer learning is not only more accurate but also more efficient.

Figure 3 shows the training and validation curves for the transfer-learning model. The loss curves drop sharply within the first few epochs and then continue to decrease slowly without any large spikes or oscillations. At the same time, both IoU and Dice increase rapidly at the beginning of training and then plateau at high values, with the validation curves closely tracking the training curves. This behavior suggests that the optimization process is stable and that the model does not suffer from severe overfitting. Compared to the Baseline and Augmented models, the transfer-learning U-Net converges faster, reaches a lower loss level, and maintains the smoothest and most stable training dynamics.

Figure 3. Transfer-learning U-Net (VGG16) training and validation loss, IoU, and Dice curves.



4.4 Baseline vs Augmented vs Transfer Learning Comparison

Figure 4 provides a direct comparison of validation behavior across the three training strategies. The Baseline model shows clear instability, with a large spike around epoch 6 and additional fluctuations afterward, confirming its lack of robustness despite strong final accuracy. The Augmented model reduces the early-stage instability seen in the Baseline curve, but still exhibits noticeable oscillations at later epochs, indicating that data augmentation improves robustness but does not fully eliminate volatility.

In contrast, the Transfer Learning model demonstrates consistently smooth and low validation loss throughout the entire training process. Its curve remains nearly flat, with no significant spikes, and stays well below the loss levels of both the Baseline and Augmented models. Together with the corresponding validation IoU and Dice curves, which are also higher and more stable across epochs, this highlights the effectiveness of leveraging pretrained VGG16 features for stable initialization, faster convergence, and better generalization. Overall, Figure 4 clearly shows that Transfer Learning achieves the most stable optimization path, while the Baseline and Augmented models suffer from varying degrees of instability.

Figure 4. Validation loss, IoU, and Dice comparison for the Baseline, Augmented, and Transfer-learning U-Net models.

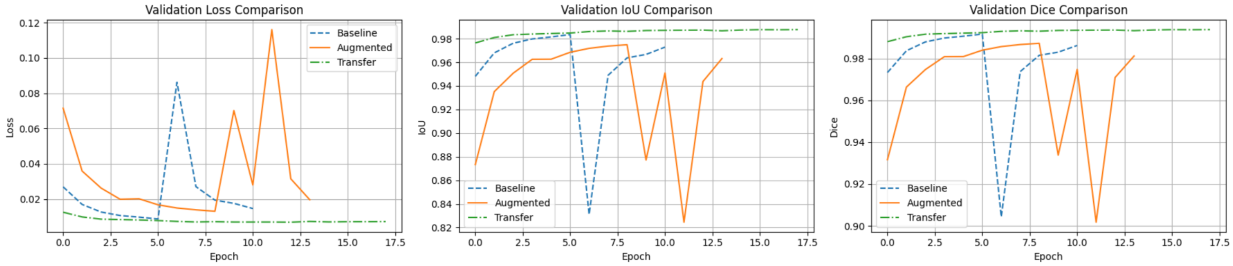


Table 1. Final test-set performance comparison.

Model	Pixel Accuracy	Loss	IoU	Dice
Baseline	0.9964	0.0091	0.9833	0.9916
Augmented	0.9945	0.0136	0.9746	0.9871
Transfer (VGG16)	0.9973	0.0071	0.9872	0.9936

The results in Table 1 further confirm that the transfer-learning model achieves the best overall performance: it has the highest pixel accuracy, IoU, and Dice score, as well as the lowest loss, outperforming both the Baseline and Augmented models on the test set.

5. Discussion

Our results highlight a fundamental trade-off between the three U-Net training strategies. The Baseline model achieves very high accuracy but suffers from catastrophic instability during training, with sharp spikes in loss and sudden drops in IoU and Dice. Data augmentation helps reduce this instability, yet it slightly lowers segmentation precision, especially around fine vehicle contours. In contrast, the transfer-learning model combines the strengths of both approaches, delivering the highest accuracy, the fastest convergence, and the most stable training dynamics.

A key finding is the impact of geometric augmentation on the Carvana dataset. The images are strictly gravity-aligned, but the augmentation pipeline applies vertical flips and random rotations, effectively creating upside-down or unusually tilted vehicles that never appear in the validation or test sets. These unrealistic views likely introduce out-of-distribution noise, forcing the model to learn invariances that are not required for the real task and weakening its ability to capture precise boundaries for standard upright vehicles.

Overall, the experiments show that while data augmentation can improve robustness for randomly initialized models, transfer learning with a pretrained VGG16 encoder is a more effective way to obtain both stability and high-quality vehicle masks.

6. Conclusion

This project demonstrates that the training strategy plays a critical role in U-Net segmentation performance. While data augmentation improves robustness and the Baseline model achieves high accuracy, the U-Net with a VGG16-based transfer-learning encoder consistently delivers the best results across all metrics. For high-precision segmentation tasks such as vehicle image masking, transfer learning represents the most effective and reliable approach among the strategies evaluated in this work.

References

Arulananth, T. S., Kuppusamy, P. G., Ayyasamy, R. K., Alhashmi, S. M., Mahalakshmi, M., Vasanth, K., & Chinnasamy, P. (2024). Semantic segmentation of urban environments: Leveraging U-Net deep learning model for cityscape image analysis. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0300767

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12)*, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Dimitrovski, I., Spasev, V., Loshkovska, S., & Kitanovski, I. (2024). U-Net ensemble for enhanced semantic segmentation in remote sensing imagery. *Remote Sensing, 16(12),* 2077. https://doi.org/10.3390/rs16122077

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., & Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems, 22(3)*, 1341–1360. https://doi.org/10.48550/arXiv.1902.07830