Do Pre-trained ImageNet Weights Enhance Road Semantic Segmentation?

Anonymous CVPR submission

Paper ID 12345

Abstract

In autonomous driving systems, road semantic segmentation plays an important role in accurately identifying road and environmental structures. This study analyzes the impact of pre-trained ImageNet weights on segmentation performance by utilizing U-Net and U-Net++ models. Three models (U-Net, U-Net++ without pre-trained weights, and U-Net++ with pre-trained weights) were compared using the KITTI dataset, and training was performed for 200 epochs. Model performance was analyzed using the Intersection-over-Union (IoU) metric and qualitative evaluation. Experimental results show that the U-Net++ model with pre-trained weights achieves the highest IoU and best captures road boundaries and object details. This study shows that transfer learning of pre-trained Imagenet dataset weights can significantly improve the performance of models that capture road structure.

1. Introduction

Semantic segmentation plays a key role in road recognition and boundary detection for autonomous vehicles. The above task of categorizing meaning on a pixel-by-pixel basis within an image is essential for applications such as autonomous vehicles. U-Net and U-Net++ are popular models in semantic segmentation, showing excellent performance in learning information about object boundaries, textures, and patterns in images. U-Net utilizes an encoder-decoder structure and skip connections to preserve high-resolution details in images. At the same time, it is designed to learn high-level contextual information. U-Net++ extends this by adding nested skip connections and dense paths, which reduce the information gap between the encoder and decoder and provide more detailed segmentation results.

Initial weight settings have a significant impact on model performance, and transfer learning is an effective way to utilize these settings. This method is effective in leveraging pre-trained weights to accelerate learning on new datasets and improve generalization performance. In the ImageNet dataset, the pre-trained weights cover a wide range of visual features. This is useful in a variety of computer vision tasks, such as semantic segmentation.

Transfer learning is a powerful tool, but it doesn't always guarantee better performance. It can actually degrade performance if there is a large difference in features between the target task and the pre-trained dataset.

In this study, we compare the road segmentation performance of U-Net, U-Net++ without pre-trained weights using VGG16 as a backbone, and U-Net++ with ImageNet pre-trained weights as initial values. The results demonstrate that pre-trained ImageNet weights are beneficial for roadway semantic segmentation. This suggests the possibility of further research utilizing different backbone models and datasets in the future.

2. Background

In this background part, we will describe the models, the dataset, and the backbone model.

2.1. Structure and differences between UNet and UNet++ models

The structure of UNet consists of a downsampling path that shrinks the input image to extract features, and an upsampling path that expands it back to restore them. Skip connections pass the output feature map from the downsampling path to the corresponding upsampling path. This design preserves high-resolution detail and makes the segmentation results precise. UNet++ extends this structure, introducing Nested Skip Connections and Nested Decoders. Each decoder block in UNet++ learns richer information by simultaneously referencing the output of multiple depths of encoders, as well as the output of the previous decoder stage. This greatly improves segmentation performance by combining different spatial resolutions.

The main difference between the two models is in their 'Dense Skip Connections' and 'Nested Architecture'. These structures are designed to reduce the semantic gap between the encoder and decoder, and to learn more fine-grained feature information. This allows UNet++ to combine feature maps of different depths, which can lead to

more sophisticated segmentation results.

2.2. Datasets Utilized

KITTI Dataset. The KITTI dataset is developed for autonomous driving research. It contains real-world road data collected by a vehicle's front-facing camera and LiDAR sensor. The dataset is suitable for addressing various challenges such as Object Detection, Semantic Segmentation, and Depth Estimation. In particular, the segmentation data in the KITTI dataset provides high-quality training data by labeling major road objects such as roads, vehicles, and pedestrians on a pixel-by-pixel basis. In this study, we utilize the segmentation labels in the KITTI dataset to train a model that more clearly distinguishes the boundaries of roads and objects. These data, when combined with the complex architecture of UNet, enable a more precise understanding of the road environment.

ImageNet Dataset. ImageNet is a large visual dataset consisting of approximately 14 million natural images and more than 1,000 classes: animals (dogs, cats, elephants), plants (trees, flowers), objects (chairs, cars), places (beaches, mountains), etc.) The dataset is designed to allow models to learn from low-level features (lines, textures, etc.) to high-level features (object shapes, etc.). Models pre-trained with ImageNet are used in a transfer learning method in a variety of computer vision tasks. By using ImageNet weights as initialization, the model starts with common visual features already learned, allowing it to learn quickly in new domains. In this study, we use VGG16 weights trained with ImageNet as initialization for the UNet++ backbone to achieve both learning efficiency and segmentation performance.

2.3. VGG16 and pre-trained weights

VGG16 is a model developed by Simonyan and Zisserman that performed well in the 2014 ImageNet Challenge. The model consists of 16 layers (13 convolutional layers and 3 fully connected layers) and is characterized by a deep structure designed using small 3x3 filters. VGG16 is pre-trained with the ImageNet dataset, which provides an effective learning state that recognizes common visual features such as edges, textures, patterns, etc. By using VGG16 as a backbone in a segmentation model, you can take advantage of the abundant visual features from the early stages of training. This can reduce training time and contribute to good performance.

3. Method

In this study, we compare the performance of three segmentation models using the augmented KITTI dataset.

1. **UNet**: Basic model trained from the beginning, without using pre-trained weights.
2. **UNet++ (no pre-trained weights)**: An extension of UNet, trained from the beginning with a model with a nested structure.
3. **UNet++ (with pre-trained weights)**: Trained using ImageNet pre-trained weights from VGG16 as initial values

The main purpose of this experiment is to evaluate how changes in model structure and the use of pre-trained weights affect segmentation performance. The reason for evaluating the UNet model together is that we needed a comparison group to determine if taking the Netsted structure always results in good performance. Each model was trained for 200 epochs. Performance metrics were evaluated periodically during the training process to determine whether the learning converged.

3.1. Training and evaluation

We used the augmented KITTI dataset for training. Augmentation techniques such as 'flipping, random cropping, and resizing' were applied to increase data diversity. During the training process, we analyzed the trend of validation loss to verify the learning stability. Specifically, we used 'IoU metrics' and 'visual evaluation' to confirm that each model was stable during learning. After training, we loaded the best performing checkpoints from each model and performed segmentation predictions on the test images. The predicted results were finally compared through 'quantitative metrics' and 'qualitative evaluation'.

3.2. Evaluation metrics

To evaluate the model performance, we used two evaluation methods.

1. **Quantitative evaluation.** The Intersection-over-Union (IoU) measures the overlap between the predicted segmentation mask and the ground truth. The formula is defined as

$$\textbf{IoU} = \frac{\textbf{Area of overlapping regions}}{\textbf{Area of union regions}}$$

   Higher IoU values indicate more accurate segmentation.

2. **Qualitative evaluation.** We visually compared the predicted masks with the Ground Truth. The Ground Truth was represented by a red mask, which allowed us to evaluate in detail how well the predictions fit.

4. Result

In this study, we evaluated three models: UNet, UNet++ (without pre-trained weights), and UNet++ (with pre-trained weights). We found that the UNet++ model with pre-trained weights performed the best.

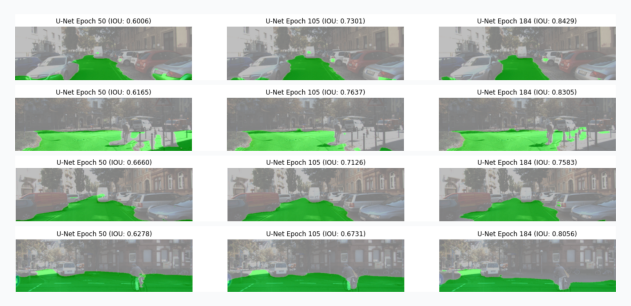

Figure 1. U-Net segmentation results at different epochs. Segmentation results for U-Net by epoch. The results show that the U-Net learns stably until the end epoch. A quantitative evaluation of the change in IOU values is also shown. We can see that the highest IOU values are seen at the end.

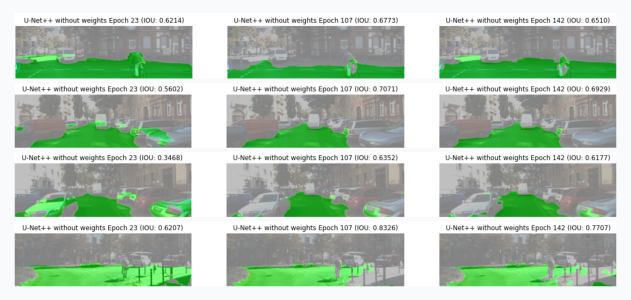

Figure 2. U-Net++ segmentation results without pre-trained weights. Unlike U-Net, we see the best quantitative and qualitative performance metrics at the midpoint of training. We can see that the checkpoint model should use a model near the midpoint above.

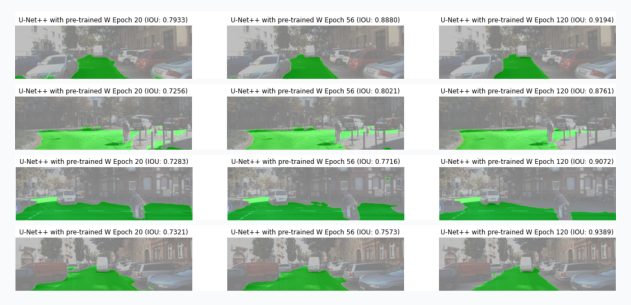

Figure 3. U-Net++ segmentation results with ImageNet pre-trained weights. This configuration demonstrates superior performance, with higher IoU values.

Figure 1-3 shows the performance of the models for each epoch using quantitative and qualitative metrics. Comparing the performance of each epoch during training provides a basis for selecting the point at which the model is optimized. The average Intersection-over-Union (IoU) value was highest for the UNet++ (using pre-trained weights) model. See Table 1 below. This means that the model did a better job of learning the ground truth regions. As a result, it segmented the road regions well.

Table 1. Average IoU values for different models.

Model	Average IoU
U-Net	0.8127
U-Net++ (No Pre-trained Weights)	0.6807
U-Net++ (Pre-trained Weights)	0.9092

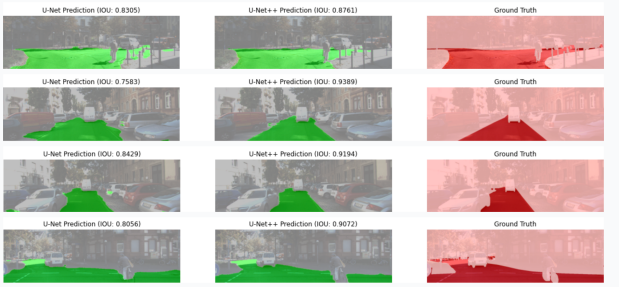

Figure 4. Comparison of predicted segmentation masks and ground truth(Red mask) for U-Net and U-Net++ with ImageNet pre-trained weights. A qualitative comparison shows that the U-Net++ model based on pre-trained weights performs better.

Figure 4 above shows a qualitative comparison of the segmentation masks of the top two performing models with the Ground Truth Mask. We can see that the UNet++ using pre-trained weights model is better at identifying the boundaries of detailed objects such as road boundaries, vehicles, and pedestrians than the original U-Net model.

5. Conclusion

These results suggest that the pre-trained weights from the ImageNet dataset were successfully transferred to a new domain such as the KITTI dataset. The ImageNet dataset consists of various classes such as animals (dogs, cats, elephants), plants (trees, flowers), objects (chairs, cars), and places (beaches, mountains). While these datasets do not directly reflect roadway segmentation, their inclusion of "transportation-related data" can provide indirect learning. In particular, they were an important contribution to learning patterns or features of background objects that should be excluded. This likely helped the model more clearly identify the edges of roads and objects during the segmentation process. The fact that the UNet++ model with pre-trained weights performed the best shows that this generalized feature learning was successfully transferred to the new domain. This once again emphasizes that the ImageNet dataset contains a wide range of visual features, and therefore has great potential to be utilized in other datasets or domains.

Additionally, the U-Net++ model (which does not use pre-

trained weights) did not clearly outperform the regular U-Net model. This suggests that U-Net++, with its nested architecture, may need to be trained on more datasets to outperform the regular U-Net model.

This study centered on the KITTI dataset, U-Net, U-Net++, and VGG16 as the backbone model. However, there are still some limitations, so the following are the directions for future research based on this study. **Diversification of backbone models.** The current study centered on the VGG16 backbone. But replacing the backbone model with 'deeper and newer models' such as ResNet and EfficientNet may further improve performance. In particular, models such as ResNet can efficiently learn from deep networks through residual connections, so we will compare their performance in future studies.

6. Acknowlegements

All of the above research was done with the help of our modulab' facilitators and peers.

7. References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pages 234–241.

[2] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pages 3–11.

[3] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. The KITTI Vision Benchmark Suite. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pages 3354–3361.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pages 248–255.

[6] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. arXiv preprint arXiv:1807.10165, 2018. URL https://arxiv.org/abs/1807.10165.

[7] GitHub Repository for UNet++. URL https://github.com/MrGiovanni/UNetPlusPlus.