UNet and Variants for Semantic Segmentation

Oday Najad
oday.najad@studenti.unipd.it

Wageesha Widuranga
wageeshawiduranga.waththeliyanage@studenti.unipd.it

Tommaso Di Tullio
tommaso.ditullio@studenti.unipd.it

January 30, 2025

Abstract

Semantic segmentation is a fundamental task in computer vision, aiming to assign a class label to each pixel in an image. This problem is particularly challenging in urban scene understanding, where complex structures and varying lighting conditions impact segmentation accuracy. Deep learning has significantly improved performance in this domain, with U-Net and its extensions being widely used due to their efficient encoder-decoder architecture.

In this project, we explore and compare the performance of three segmentation models: U-Net, Nested U-Net (U-Net++), and Attention U-Net. These architectures are evaluated on an urban street dataset with pixel-wise annotations. The models are trained using a combination of Cross-Entropy, Dice, and IoU losses to optimize segmentation performance.

Our experiments show that while the standard U-Net provides a strong baseline, Nested U-Net enhances feature propagation through dense skip connections, and Attention U-Net further improves segmentation by leveraging attention mechanisms. The results indicate that Attention U-Net achieves superior segmentation accuracy, particularly in handling occlusions and fine details. The findings highlight the strengths and limitations of each model and provide insights into their applicability for real-world semantic segmentation tasks.

1 Introduction

Deep learning has significantly advanced semantic segmentation, with convolutional neural networks (CNNs) playing a crucial role. Among the most effective architectures, U-Net [Ronneberger et al., 2015] and its variants have been widely adopted due to their encoder-decoder structure and skip connections, which enhance spatial information retention. Variants like Nested U-Net (U-Net++) [Zhou et al., 2018] introduce dense skip pathways to refine feature propagation, while Attention U-Net [Oktay et al., 2018] incorporates attention mechanisms to focus on relevant image regions, improving segmentation accuracy.

In this project, we explore and compare these architectures for urban scene segmentation. We train and evaluate U-Net, Nested U-Net, and Attention U-Net on a benchmark dataset, analyzing their performance in terms of segmentation accuracy, computational efficiency, and robustness to occlusions. The models are trained using a combination of Cross-Entropy, Dice, and IoU losses to optimize both pixel-wise classification and boundary accuracy.

2 Related Work

2.1 U-Net in Semantic Segmentation

U-Net [Ronneberger et al., 2015] was originally introduced for biomedical image segmentation and has since been widely adopted for various segmentation tasks. The architecture follows an encoder-decoder structure, where the contracting path captures context, and the symmetric expansive path enables precise localization.

The effectiveness of U-Net in semantic segmentation has been demonstrated across multiple domains. For example, in medical imaging, it has been successfully applied for segmenting organs and tumors in CT and MRI scans [Çiçek et al., 2016, Milletari et al., 2016]. Similarly, in remote sensing, U-Net has been used for land cover classification and urban scene segmentation [Mou & Zhu, 2019, Zhang et al., 2018]. Furthermore, its extension to 3D U-Net has proven beneficial in volumetric segmentation tasks [Çiçek et al., 2016].

Several improvements to the original U-Net have been proposed, such as the use of residual connections [Drozdzal et al., 2016] and densely connected architectures [Guan et al., 2019], which help in mitigating the vanishing gradient problem and improving feature reuse.

2.2 Nested U-Net (U-Net++) in Semantic Segmentation

U-Net++ [Zhou et al., 2018] enhances the original U-Net by introducing dense skip pathways, designed to bridge the semantic gap between encoder and decoder representations. This architecture refines segmentation accuracy by ensuring better information flow across different levels of abstraction.

The introduction of U-Net++ has led to significant performance gains in medical imaging [Zhang et al., 2019]. Studies have shown that it outperforms standard U-Net in tasks such as lung nodule segmentation [Wang et al., 2020] and brain tumor segmentation [Isensee et al., 2018]. The dense connections in U-Net++ facilitate better feature propagation, leading to more refined segmentation maps.

In addition to medical imaging, U-Net++ has also been utilized in satellite image analysis and urban planning applications [Ji et al., 2019, Liu et al., 2020], where fine-grained segmentation is crucial. The architecture's robustness in handling complex structures makes it a preferred choice for high-resolution image segmentation.

2.3 Attention U-Net in Semantic Segmentation

Attention U-Net [Oktay et al., 2018] extends U-Net by incorporating attention mechanisms, which allow the model to focus on the most relevant features in an image. The attention gates suppress irrelevant background information while enhancing feature representation in critical regions.

This model has demonstrated superior performance in medical imaging, particularly in applications such as cardiac MRI segmentation and liver tumor detection [Schlemper et al., 2019]. By dynamically weighting spatial features, Attention U-Net improves segmentation accuracy while reducing computational overhead.

Beyond medical imaging, Attention U-Net has been applied in remote sensing and environmental monitoring, where precise object delineation is required [Zhang et al., 2021, Ma et al., 2020]. The inclusion of attention modules enables the network to differentiate between similar-looking objects, making it effective in segmenting roads, buildings, and natural landscapes.

3 Dataset

3.1 Dataset Overview

The Cityscapes dataset consists of **5,000 finely annotated** images and **20,000 coarsely annotated** images, captured from **50 different cities**. These images cover a wide variety of urban environments, traffic conditions, and weather scenarios, making it an ideal dataset for training robust semantic segmentation models.

Each image is provided in **high resolution (2048×1024 pixels)** and contains **30 object classes**, which are grouped into **8 supercategories**: **Flat surfaces:** road, sidewalk, parking, rail track. **Human-made structures:** building, wall, fence, bridge. **Nature:** vegetation, terrain. **Sky:** sky. **Vehicles:** car, truck, bus, train, motorcycle, bicycle. **People:** person, rider. **Objects:** pole, traffic sign, traffic light. **Void:** unlabelled regions and ambiguous areas.

3.2 Annotations and Masks

Each image in the dataset is accompanied by its corresponding **semantic segmentation mask**, where each pixel is labeled with its respective class. The masks use a color-coded scheme, making it easy to visually distinguish between different object categories.

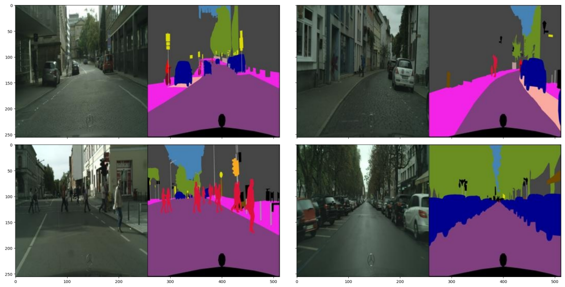

Figure 1: Examples from the Cityscapes dataset.

3.3 Training and Validation Split

For training and evaluation purposes, we used the **official split** provided by Cityscapes(training set: 2,975 finely annotated images, validation set: 500 finely annotated images, test set: 1,525 finely annotated images (without public ground truth)).

3.4 Data Preprocessing

Before feeding the dataset into our models, several preprocessing steps were applied to ensure consistency and improve training efficiency:

- **Resizing and Normalization**: All input images were resized to a fixed resolution suitable for training while pixel values were normalized to range between $[0, 1]$.

- **Categorical Label Mapping**: The segmentation masks were originally stored as RGB images, where different colors represented different object categories. These masks were converted into class indices to simplify the learning process.

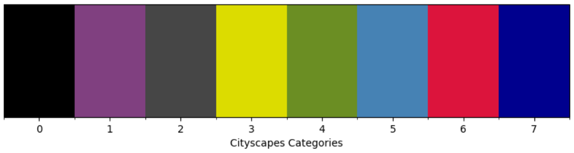

Figure 2: Visualization of Cityscapes categories with corresponding colors.

- **One-Hot Encoding**: Since the segmentation task involves multiple categories, each label was transformed into a one-hot encoded tensor, making it easier for the model to learn per-pixel class probabilities.

- **Data Augmentation**: To increase dataset variability and improve model generalization, transformations such as random cropping, horizontal flipping, brightness adjustment, and contrast normalization were applied.

- **Conversion to Tensor Format**: Both input images and segmentation masks were converted to tensor format, ensuring compatibility with deep learning frameworks.

4 Method

4.1 U-Net

U-Net is a fully convolutional network (FCN) designed for biomedical image segmentation, originally introduced by Ronneberger et al. [Ronneberger et al., 2015]. It follows an encoder-decoder architecture with skip connections that allow the network to capture both global context and fine details. The network is structured symmetrically, consisting of a contracting path (encoder) and an expansive path (decoder).

The encoder progressively downsamples the input image through convolutional layers followed by max-pooling operations. The feature maps are then upsampled in the decoder using transposed convolutions, restoring the spatial resolution. Skip connections concatenate feature maps from the encoder to corresponding layers in the decoder to recover lost spatial information.

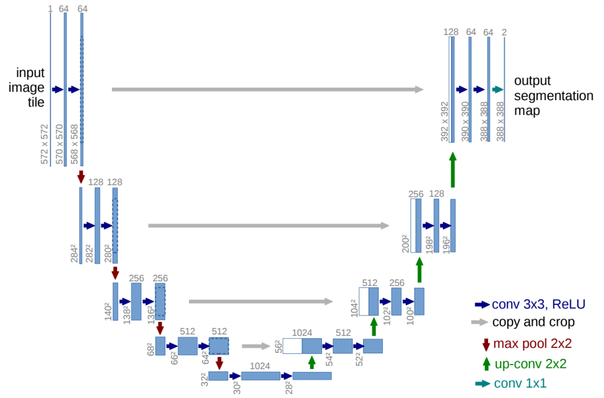

Figure 3: U-Net architecture adapted from Ronneberger et al. [Ronneberger et al., 2015]. The network consists of an encoder (left side), decoder (right side), and skip connections.

Mathematical Formulation The key operation in U-Net is the convolutional layer, defined as:

$$f(x) = \sigma(W * x + b) \tag{1}$$

where: x is the input tensor, W is the learned weight kernel, $*$ denotes the convolution operation, b is the bias term, σ represents the activation function (ReLU in U-Net).

The downsampling step involves max pooling:

$$y_{i,j} = \max_{(m,n)\in\mathcal{R}(i,j)} x_{m,n} \tag{2}$$

where $\mathcal{R}(i,j)$ denotes the receptive field around pixel (i,j).

Upsampling in the decoder uses transposed convolution:

$$y = W^T * x \tag{3}$$

which increases the spatial resolution while learning new feature representations.

Finally, the network outputs a segmentation mask $\hat{Y}$, computed as:

$$\hat{Y} = \text{Softmax}(W_{\text{out}} * x_{\text{final}} + b_{\text{out}}) \tag{4}$$

where Softmax ensures that the output at each pixel corresponds to class probabilities.

4.2 Nested U-Net (U-Net++)

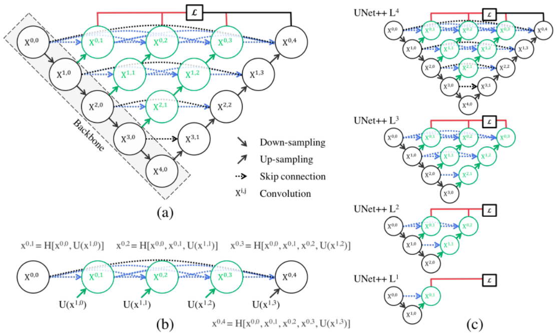

Figure 4: U-Net++ architecture adapted from Zhou et al. [Zhou et al., 2018]. The network consists of densely connected convolutional layers, deep supervision, and progressive feature refinement.

U-Net++ is an improved extension of the U-Net architecture, designed to enhance feature fusion through dense convolutional blocks and deep supervision [Zhou et al., 2018]. The main goal of U-Net++ is to bridge the semantic gap between the encoder and decoder by introducing a nested and dense skip connection structure, as illustrated in Figure 4. Unlike the original U-Net, which employs simple skip connections, U-Net++ incorporates a series of dense convolutional blocks between the encoder and decoder paths. The network consists of multiple convolutional layers at varying depths, where each intermediate layer is progressively refined before being passed to the next level.

Let $x^{i,j}$ represent the feature map at the j^{th} convolutional layer in the i^{th} level of the encoder. The feature propagation is defined as:

$$x^{i,j} = H\left(x^{i,0}, x^{i-1,j}, U(x^{i-1,j})\right) \tag{5}$$

where: $H(\cdot)$ represents a series of convolutional and activation functions, $x^{i-1,j}$ is the feature map from the previous depth level, $U(\cdot)$ denotes an upsampling operation that ensures spatial alignment.

The nested structure ensures that each decoder layer is progressively refined before final segmentation.

4.3 Attention U-Net

Attention U-Net is an enhancement of the standard U-Net architecture that integrates an **attention mechanism** to dynamically highlight salient features while suppressing irrelevant regions [Oktay et al., 2018].

Attention Mechanism The key idea behind Attention U-Net is to learn **spatial attention maps**. The attention gating mechanism is formally defined as:

$$\alpha = \sigma\left(W_g^T g + W_x^T x + b\right) \tag{6}$$

The computed attention coefficient α is then applied to the input feature map x:

$$\tilde{x} = \alpha \odot x \tag{7}$$

where $\odot$ represents element-wise multiplication. This mechanism ensures that the network focuses on the most relevant regions while suppressing background noise, leading to improved segmentation accuracy.

$$\alpha = \sigma_2\left(\psi^T\left(\sigma_1\left(W_g g + W_x x^l\right)\right)\right) \tag{8}$$

where: g is the input from the **decoder**, x^l is the input from the **encoder skip connection**, W_g and W_x are 1×1 convolutions that match feature dimensions, ψ^T is another 1×1 convolution, σ_1 is the **ReLU** activation function, σ_2 is the **sigmoid** activation function, α represents the learned attention coefficients.

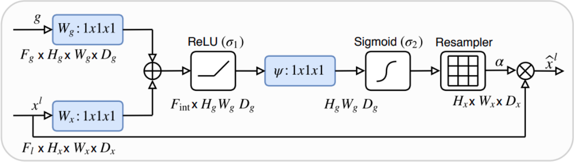
g
$W_g: 1x1x1$
$F_g \times H_g \times W_g \times D_g$
x^l
$W_x: 1x1x1$
$F_l \times H_x \times W_x \times D_x$
ReLU (σ_1)
$F_{int} \times H_g W_g D_g$
$\psi: 1x1x1$
Sigmoid (σ_2)
$H_g W_g D_g$
Resampler
α
$H_x \times W_x \times D_x$
$\hat{x}^l$

Figure 5: Attention mechanism in Attention U-Net [Oktay et al., 2018].

Feature Reweighting The attention coefficients α modulate the encoder features before being passed to the decoder:

$$\hat{x}^l = \alpha \cdot x^l \tag{9}$$

This mechanism ensures that only the most relevant encoder features are propagated, improving the network's ability to distinguish between background and foreground objects.

5 Results

5.1 Training and Convergence Analysis

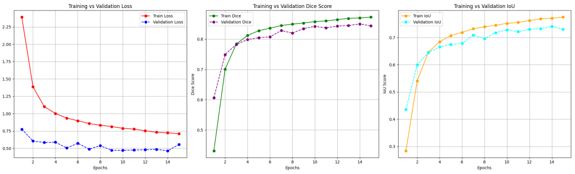

Figure 6: Training performance of U-Net.

U-Net: The loss curve shows a steady decline, with the Dice Score and IoU improving over epochs. However, the validation loss plateaus earlier than the more advanced architectures, suggesting limited capacity to generalize.

Nested U-Net (U-Net++): The introduction of nested skip connections facilitates better gradient flow, leading to a more stable and smoother convergence curve. The model achieves slightly higher Dice and IoU scores than standard U-Net.

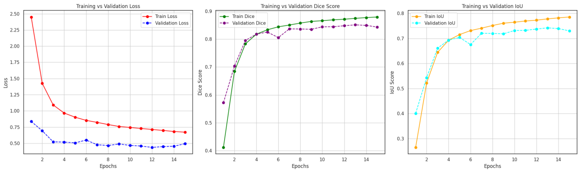

Figure 7: Training performance of Nested U-Net (U-Net++).

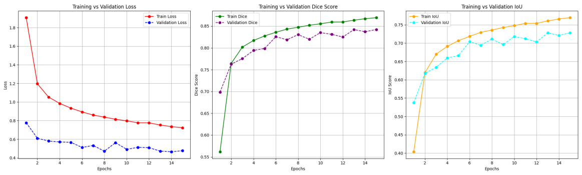

Figure 8: Training performance of Attention U-Net.

Attention U-Net: This architecture consistently outperforms the others in both training and validation metrics. The attention mechanism enables the model to focus on relevant regions of the image, resulting in improved segmentation quality. Notably, the validation Dice Score stabilizes at a higher value compared to the other models.

Table 1: Comparison of U-Net, U-Net++, and Attention U-Net performance.

Model	**Dice Score**	**IoU Score**
U-Net	0.851	0.742
Nested U-Net	0.857	0.750
Attention U-Net	0.864	0.758

Table 2: Comparison with State-of-the-Art Segmentation Models.

Model	**Dice Score**	**IoU**
DeepLabV3+	0.89	0.79
PSPNet	0.88	0.78
U-Net	0.851	0.742
Nested U-Net	0.857	0.750
Attention U-Net	0.864	0.758

5.2 Comparison with State-of-the-Art Results

6 Conclusion

In this study, we explored the effectiveness of U-Net and its advanced variants, Nested U-Net (U-Net++) and Attention U-Net, for semantic segmentation on the Cityscapes dataset. Our analysis covered architectural differences, training performance, and segmentation accuracy, offering a comprehensive comparison of these models.

The results demonstrate that while the standard U-Net provides a strong baseline for semantic segmentation, improvements in architectural design significantly enhance its performance. **Nested U-Net (U-Net++)** leverages dense skip connections and deep supervision to refine segmentation boundaries, leading to better feature propagation. **Attention U-Net** further improves results by incorporating attention mechanisms that focus on the most relevant parts of the image, resulting in superior segmentation accuracy, particularly for small and complex objects.

Despite these improvements, our models achieved slightly lower performance compared to state-of-the-art segmentation frameworks such as DeepLabV3+ and PSPNet. The gap in performance can be attributed to the absence of atrous convolutions, pretrained backbones, and multi-scale feature extraction techniques commonly used in top-tier models. However, our findings reinforce the adaptability and efficiency of U-Net-based architectures for real-world semantic segmentation tasks.

References

[Drozdzal et al., 2016] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., & Pal, C. (2016). The importance of skip connections in biomedical image segmentation. *Deep Learning and Data Labeling for Medical Applications*, (pp. 179–187).

[Guan et al., 2019] Guan, S., Khan, A., Sikdar, S., & Chitnis, P. V. (2019). Fully dense unet for 2d sparse photoacoustic tomography artifact removal. *IEEE Journal of Biomedical and Health Informatics*, 23(2), 642–650.

[Isensee et al., 2018] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2018). nnu-net: a self-adapting framework for u-net-based medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 3–11).: Springer.

[Ji et al., 2019] Ji, S., Wu, Y., Xie, W., & Zhang, Y. (2019). Improved unet++ architecture for remote sensing image segmentation. *IEEE Access*, 7, 99512–99522.

[Liu et al., 2020] Liu, H., Zhou, C., Zou, Z., Yu, Q., Wang, Z., & Jiang, X. (2020). Road segmentation based on deep learning from high-resolution remote sensing imagery. *Remote Sensing*, 12(14), 2231.

[Ma et al., 2020] Ma, L., Zhou, X., Li, W., et al. (2020). Hyper-spectral image classification using attention u-net. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8), 5523–5534.

[Milletari et al., 2016] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *3D Vision (3DV), 2016 Fourth International Conference on*, (pp. 565–571).

[Mou & Zhu, 2019] Mou, L. & Zhu, X. X. (2019). Relation matters: Foreground-aware relational reasoning for human-object interaction detection. *IEEE Transactions on Geoscience and Remote Sensing*.

[Oktay et al., 2018] Oktay, O., Schlemper, J., Folgoc, L. L., et al. (2018). Attention u-net: Learning where to look for the pancreas. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[Schlemper et al., 2019] Schlemper, J., Oktay, O., Schaap, M., et al. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53, 197–207.

[Wang et al., 2020] Wang, X., Ma, H., Song, X., & Huang, X. (2020). Automated lung nodule segmentation using improved u-net++. *IEEE Access*, 8, 123951–123959.

[Zhang et al., 2018] Zhang, J., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5), 749–753.

[Zhang et al., 2021] Zhang, L., Zhang, B., Tang, H., & Ma, A. (2021). Attention-based semantic segmentation of high-resolution remote sensing images. *Remote Sensing*, 13(8), 1512.

[Zhang et al., 2019] Zhang, Z., Liu, Q., Wang, Y., & Chang, E. I.-C. (2019). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 16(5), 740–744.

[Zhou et al., 2018] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.

[Çiçek et al., 2016] Çiçek, , Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3d u-net: Learning dense volumetric segmentation from sparse annotation. (pp. 424–432).