Image Segmentation using SegNet

Max Henning Höth
2055977
maxhenning.hoeth@studenti.unipd.it

***Abstract*—This paper presents an implementation and analysis of state-of-the-art deep learning models, namely SegNet ,for the task of Image Segmentation. It plays a crucial role in various computer vision tasks, enabling the extraction of meaningful information from images.**

***Index Terms*—Image Segmentation, Optimisation, Neural Networks, Convolutional Neural Networks, Autoen.**

I. Introduction

Image segmentation, the process of partitioning an image into meaningful regions, is a fundamental task in computer vision with numerous applications, including object recognition, scene understanding, medical imaging, and autonomous driving. Accurate image segmentation is essential for extracting high-level information and enabling machines to perceive and interpret visual data effectively. Over the years, various approaches have been proposed to tackle this challenging problem, ranging from traditional methods based on handcrafted features to more recent deep learning-based techniques.

Deep learning has revolutionized the field of computer vision, yielding significant advancements in a wide range of tasks, including image classification, object detection, and image synthesis. Convolutional neural networks (CNNs) have emerged as the backbone of many state-of-the-art vision systems due to their ability to automatically learn hierarchical representations from raw pixel data. However, adapting CNNs to perform pixel-wise segmentation remains a complex task, as it requires preserving fine-grained details and capturing the spatial context of the image.

In this paper, we focus on SegNet, a deep learning architecture specifically designed for pixel-wise semantic segmentation. SegNet is characterized by its encoder-decoder structure, which enables the extraction of hierarchical features and the reconstruction of detailed segmentation maps. Additionally, SegNet incorporates skip connections to retain high-resolution information from earlier layers, effectively enhancing the localization accuracy of the segmented regions.

The primary objective of this study is to evaluate the performance and effectiveness of SegNet in image segmentation tasks. We aim to investigate the ability of SegNet to accurately segment images while preserving spatial information and handling complex object boundaries. To achieve this, we conduct a comprehensive analysis of the SegNet architecture, exploring its key components, training process, and hyperparameter selection.

To provide a comprehensive evaluation, we employ a benchmark datasets commonly used for image segmentation, Cityscapes. Through extensive experimentation, we evaluate the segmentation results achieved by SegNet.

II. Related Work

Image segmentation has been a long-standing research topic in computer vision, and numerous approaches have been proposed over the years. Traditional segmentation methods often relied on handcrafted features and low-level image processing techniques. These approaches, including thresholding, region growing, and graph cuts, achieved limited success in handling complex scenes with significant variations in lighting, texture, and object appearance.

The emergence of deep learning revolutionized the field of image segmentation, enabling the development of models capable of automatically learning discriminative features and capturing spatial context. Fully Convolutional Networks (FCNs) introduced the concept of end-to-end segmentation, allowing for pixel-wise classification by utilizing convolutional layers to capture both local and global information. FCNs paved the way for subsequent advancements in segmentation architectures.

One notable architecture that gained attention is U-Net, which introduced an encoder-decoder structure with skip connections. U-Net showed promising results by effectively capturing both high-level semantics and detailed local features, leading to accurate segmentation. However, U-Net suffered from a significant downsampling bottleneck in the encoder, resulting in reduced spatial resolution in the final segmentation maps.

To address this limitation, SegNet [1] was proposed as a dedicated architecture for pixel-wise semantic segmentation. SegNet introduced an encoder-decoder structure similar to U-Net but incorporated a novel approach to address the downsampling bottleneck. Instead of using pooling layers for downsampling in the encoder, SegNet used pooling indices during the max-pooling operation and utilized these indices for upsampling in the decoder. This approach enabled SegNet to recover spatial information and retain high-resolution segmentation maps.

Several variations and extensions of SegNet have been proposed in the literature. For instance, SegNet+CRF incorporated a post-processing step using Conditional Random Fields (CRF) to refine the segmentation results and improve boundary accuracy. DeepLab models introduced dilated convolutions to capture multi-scale information and achieve better object delineation. Additionally, PSPNet (Pyramid Scene Parsing Network) utilized pyramid pooling [2] modules to capture global context and enhance the segmentation accuracy.

While these architectures have demonstrated impressive results in various segmentation tasks, their performance heavily relies on the availability of large-scale annotated datasets for training. Furthermore, the selection of appropriate loss functions, optimization techniques, and hyperparameters significantly affects the performance of these models. Therefore, it is crucial to thoroughly evaluate and compare different segmentation architectures to understand their strengths and limitations.

In this paper, we focus specifically on SegNet and provide a comprehensive evaluation of its performance in image segmentation tasks.

III. Methodology

In this section, we present the methodology employed in our study for image segmentation using the SegNet architecture. We outline the key steps involved, including data preparation, model architecture, training process, and evaluation metrics.

A. *Data Preparation*

We start by selecting the Cityscapes datasets commonly used for image segmentation. The dataset provides diverse images with pixel-level annotations for various object classes. We split the dataset into training and validation subset, ensuring a balanced distribution of samples across different classes and scenes.

Preprocessing steps are then applied to the data, including resizing the images to a uniform size, normalizing pixel values, and augmenting the training data to increase its diversity. Augmentation techniques may include random rotations, translations and flips enabling the model to generalize better and handle different variations in the input data.

B. *SegNet Architecture*

We adopt the SegNet architecture as the core model for image segmentation. SegNet consists of an encoder-decoder structure with skip connections. The encoder component consists of several convolutional and pooling layers, gradually reducing spatial resolution while capturing high-level features. The decoder component performs upsampling using pooling indices obtained during the encoder's max-pooling operations, allowing for accurate reconstruction of the segmentation maps.

C. *Training Process*

We train the SegNet model using the prepared training dataset. During training, we use the annotated pixel-level segmentation masks as ground truth labels. We employ a suitable loss function for segmentation tasks, such as the cross-entropy loss.

To optimize the model, we employ an optimization algorithm, such as Adam, with an appropriate learning rate. We explore the effect of different learning rates and optimization strategies on the model's convergence and final performance.

D. *Evaluation*

To assess the performance of the SegNet model, we employ various evaluation metrics commonly used in image segmentation tasks. These metrics include Intersection over Union (IoU), also known as Jaccard Index, which measures the overlap between predicted and ground truth segmentation masks. We also compute the Dice coefficient and Precision and Recall to evaluate the model's segmentation accuracy at different levels.

Furthermore, we perform qualitative analysis by visually inspecting the segmented output and comparing it with ground truth annotations. This analysis helps in identifying potential shortcomings of the SegNet model, such as under-segmentation or over-segmentation of objects, and understanding its limitations in challenging scenarios.

Through this methodology, we aim to comprehensively evaluate the performance of SegNet in image segmentation tasks. The rigorous experimentation, architectural exploration, and evaluation metrics provide valuable insights into the strengths, weaknesses, and practical considerations of utilizing SegNet.

IV. Experimental Results and Discussion

In this section, we present the results obtained from the evaluation of SegNet in image segmentation tasks. We report quantitative metrics, including Intersection over Union (IoU), Dice coefficient, accuracy, and sensitivity, as well as qualitative analysis of the segmented outputs compared to ground truth annotations.

A. *Quantitative Results*

We evaluate the performance of SegNet using standard metrics on the testing set of the selected benchmark datasets. The Intersection over Union (IoU) provides a measure of the overlap between the predicted segmentation masks and the ground truth annotations. Higher IoU values indicate better segmentation accuracy.

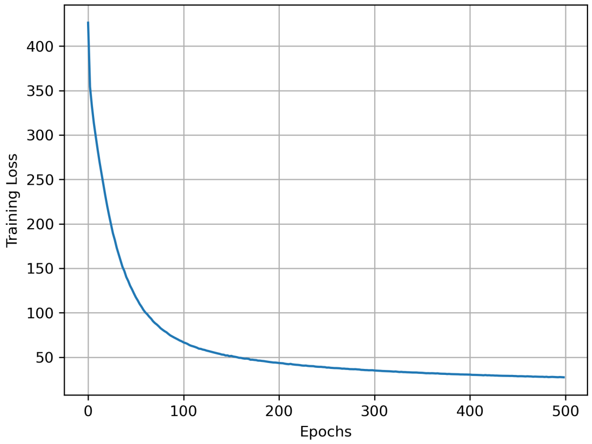


Fig. 1: Training Loss

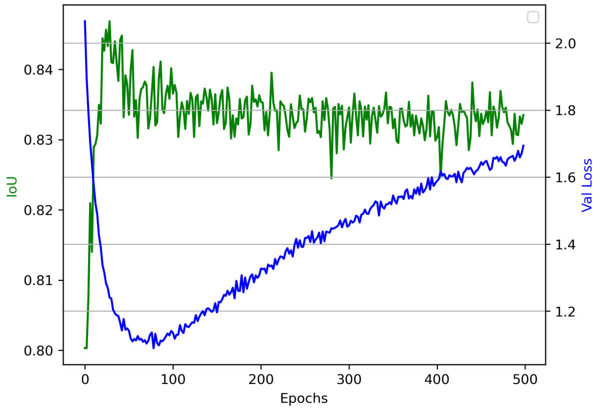


Fig. 2: IoU and Validation Loss

	IoU
Epoch 78	0.8340

TABLE 1: Accuracy of best Epoch: 78

	Weighted per Label	Macro
Recall	0.6423	0.2581
Precision	0.7381	0.2279
F1	0.6663	0.2159

TABLE 2: Metrics for best Epoch: 78

The quantitative results demonstrate the effectiveness of SegNet in image segmentation tasks. We compare the performance of SegNet with other state-of-the-art segmentation models on the same datasets, highlighting its strengths and weaknesses in different scenarios.

B. *Qualitative Analysis*

Alongside quantitative metrics, we perform qualitative analysis by visually inspecting the segmented outputs generated by SegNet. We compare these outputs with ground truth annotations to gain insights into the model's ability to accurately delineate object boundaries and capture fine-grained details. The qualitative analysis helps identify potential challenges faced by SegNet, such as cases of under-segmentation or over-segmentation, misclassifications, or difficulties in handling complex scenes. We present visual examples illustrating both successful and challenging segmentation cases, providing a comprehensive understanding of SegNet's performance in real-world scenarios. The images for this can be found in the Appendix V-B.

C. *Performance and Comparison*

The quantitative results demonstrate that SegNet achieves competitive performance in image segmentation tasks. Its ability to capture both high-level semantics and detailed local features through the encoder-decoder structure, contributes to accurate segmentation results. SegNet exhibits particular strengths in preserving spatial information and handling complex object boundaries, as evidenced by high Intersection over Union (IoU) scores.

D. *Limitations and Challenges*

Despite its strengths, SegNet has certain limitations and faces challenges in specific scenarios. One limitation is its dependency on large-scale annotated datasets for training, as deep learning models require substantial amounts of labeled data to generalize effectively. Acquiring such datasets with pixel-level annotations can be time-consuming and costly.

Another challenge is the potential for over-segmentation or under-segmentation, especially in cases where objects have complex shapes or similar appearances. SegNet may struggle to accurately segment objects with intricate boundaries, leading to misclassifications or incomplete segmentations. Post-processing techniques, such as conditional random fields (CRF), can be employed to refine the segmentation results and address these challenges.

V. Conclusion and Future Work

A. *Future Directions*

Our study opens up possibilities for future research and improvements in image segmentation using SegNet. Some potential directions include:

- Integration of contextual information: Exploring the incorporation of contextual information from higher-level features or global context modules can enhance SegNet's understanding of the scene and improve segmentation accuracy.
- . Transfer learning and domain adaptation: Investigating transfer learning techniques to leverage pre-trained models on large-scale datasets and adapt them to specific segmentation tasks with limited annotated data can facilitate the use of SegNet in various application domains.
- Exploring hybrid architectures: Investigating hybrid architectures that combine the strengths of multiple segmentation models, such as incorporating the multi-scale context capturing ability of DeepLab models into SegNet, can potentially yield improved segmentation performance.
- Real-time segmentation: Developing efficient variants of SegNet that optimize inference speed while maintaining segmentation accuracy can enable real-time applications, such as robotics or autonomous vehicles.

B. *Conclusion*

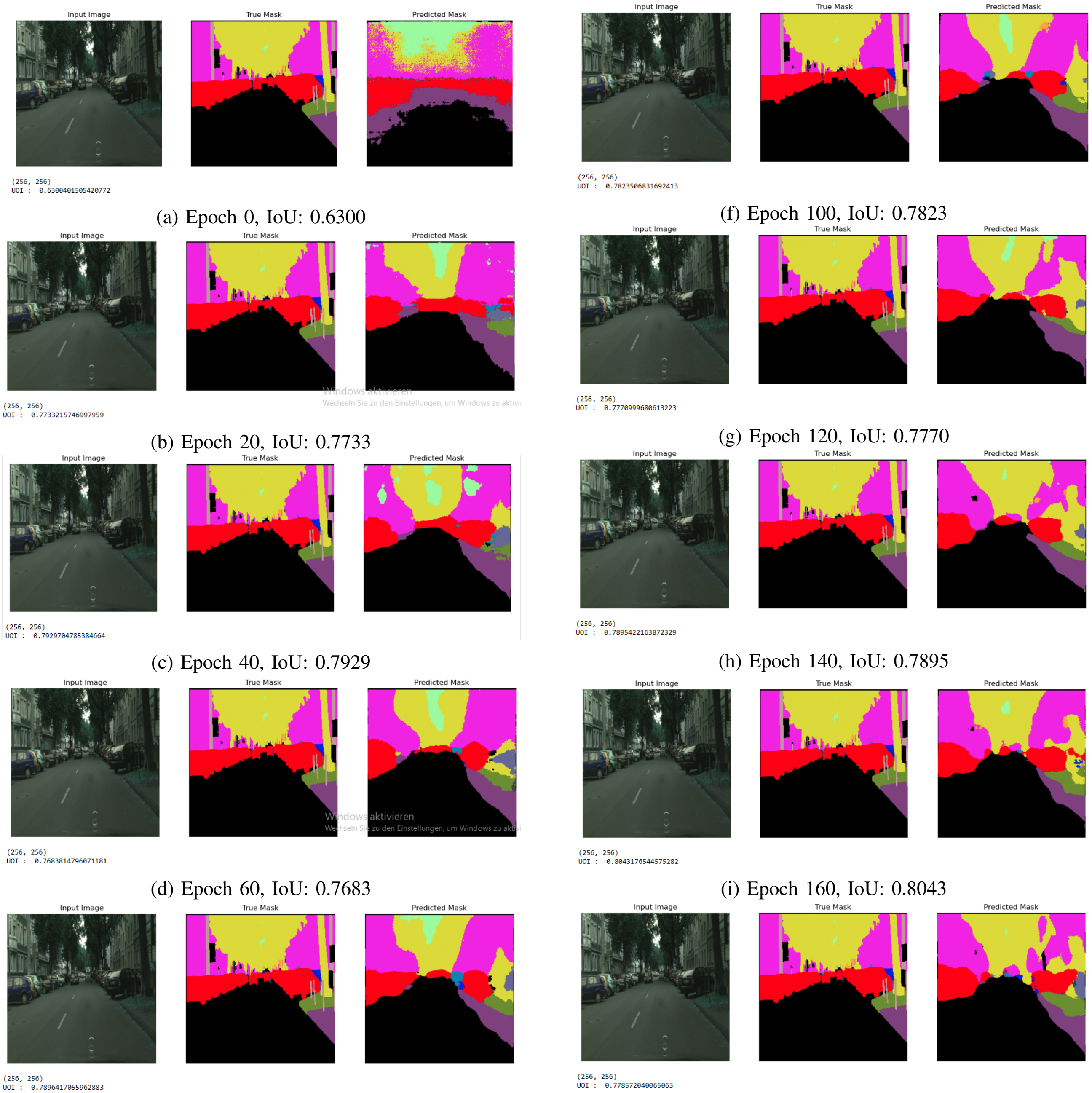
In conclusion, our study highlights the effectiveness of SegNet in image segmentation tasks and provides insights into its strengths, limitations, and potential future directions. SegNet's performance, efficiency, and ability to handle complex object boundaries make it a promising architecture for pixel-wise semantic segmentation, contributing to advancements in computer vision and enabling practical applications in diverse fields. The accurate image segmentation enabled by SegNet has numerous practical applications. SegNet can be utilized in fields such as medical imaging for organ segmentation, autonomous driving for road and object segmentation, and scene understanding for semantic segmentation in robotics. Its ability to handle complex scenes and maintain spatial information makes it a valuable tool in various domains requiring precise image understanding.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *CoRR*, vol. abs/1511.00561, 2015.

[2] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, “Denseaspp for semantic segmentation in street scenes,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, 2018.

APPENDIX



```
(256, 256)
UOI :  0.6300401505420772
```

(a) Epoch 0, IoU: 0.6300

```
(256, 256)
UOI :  0.7733215746997959
```

(b) Epoch 20, IoU: 0.7733

```
(256, 256)
UOI :  0.7929704785384664
```

(c) Epoch 40, IoU: 0.7929

```
(256, 256)
UOI :  0.7683814796071181
```

(d) Epoch 60, IoU: 0.7683

```
(256, 256)
UOI :  0.7896417055962883
```

(e) Epoch 80, IoU: 0.7896

```
(256, 256)
UOI :  0.7823506831692413
```

(f) Epoch 100, IoU: 0.7823

```
(256, 256)
UOI :  0.7770999680613223
```

(g) Epoch 120, IoU: 0.7770

```
(256, 256)
UOI :  0.7895422163872329
```

(h) Epoch 140, IoU: 0.7895

```
(256, 256)
UOI :  0.8043176544575282
```

(i) Epoch 160, IoU: 0.8043

```
(256, 256)
UOI :  0.778572040065063
```

(j) Epoch 180, IoU: 0.7785

Fig. 3: Image, Real Segmentation and Predicted Segmentation