Semantic Segmentation Using the U-Net Architecture on Monocular Datasets

Ahmad Fikri Hanafi[1], Ervin Yohannes[2]

[1,2] Informatics Engineering / Bachelor Program of Informatics Engineering, State University of Surabaya
[1]ahmad.21074@mhs.unesa.ac.id
[2]ervinyohannes@unesa.ac.id

***Abstract*— This study implements a deep learning model based on the U-Net architecture with a pre-trained ResNet50 backbone on ImageNet to solve the task of semantic segmentation on monocular images. The Cityscapes dataset is used as the main benchmark because it provides high-quality data with high resolution that is widely recognized in urban image segmentation research. Experiments were conducted to evaluate the model's performance with varying learning rate values, aiming to understand the model's sensitivity to training parameters. The results show that a learning rate of 1e-4 yields optimal performance, achieving a Mean Intersection over Union (Mean IoU) of 86.59% and pixel accuracy of 97.63%. Visualization of the segmentation predictions demonstrates the model's ability to accurately recognize urban objects and structures, especially under varying lighting conditions and background complexity. These findings confirm the effectiveness of U-Net in image segmentation tasks, as well as the importance of hyperparameter selection and dataset quality in achieving high model performance in the monocular image domain.**

***Keywords*— *Convolusional Neural Network*, *Deep Learning*, U-Net, *Encoder-Decoder*, Semantic Segmentation.**

I. INTRODUCTION

Advances in artificial intelligence (AI) technology and deep learning models have led to significant developments in many areas, one of which is digital image processing. One application that is increasingly being applied is image segmentation in infrastructure mapping, such as monocular dataset analysis. Monocular dataset segmentation plays a crucial role in various applications, including autonomous navigation systems, traffic monitoring, and urban planning [1].

However, monocular dataset segmentation from aerial or ground images faces various challenges, such as inadequate lighting, shadows, vehicles, and complex surrounding environments [2]. In recent years, deep learning models based on Convolutional Neural Networks (CNNs) have demonstrated highly significant and accurate performance in image segmentation tasks compared to traditional methods [3]. One of the most well-known CNN architectures for image segmentation is U-Net. U-Net was originally developed for medical image segmentation but has been successfully adapted to various domains, including monocular dataset analysis [4]. Research shows that the U-Net architecture can perform image segmentation tasks with satisfactory results, even on high-resolution images [5].

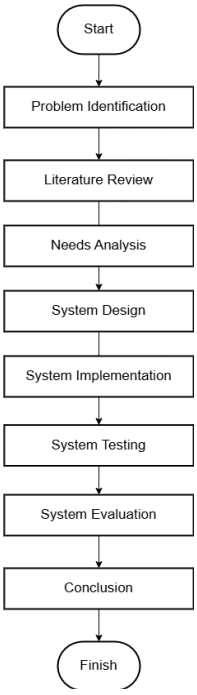
Recent research indicates that combining U-Net with appropriate data augmentation and preprocessing techniques can improve segmentation performance under various image conditions [2]. However, the main challenge in implementing U-Net is the high computational requirements, especially when training models using large monocular datasets. The use of model optimization techniques such as transfer learning and learning rate optimizers is crucial for improving efficiency and reducing training time. The application of these techniques can help reduce high computational resource requirements without compromising segmentation quality [4].

The quality of training image data is a critical aspect in the implementation of monocular dataset segmentation models. Datasets with accurate annotations and representative of various conditions in monocular datasets will significantly influence the results obtained [3]. Issues in datasets, such as limited environmental variation and poor lighting, can cause models to fail to recognize specific environments. Therefore, a semi-supervised approach or the addition of synthetic datasets can be a solution to improve model accuracy [1]

Although the U-Net model has proven its effectiveness in monocular dataset segmentation, several challenges remain. These challenges include the model's need to recognize various conditions in monocular datasets with limited data, as well as parameter and model architecture adjustments to overcome constraints such as the presence of complex objects, diverse backgrounds, and varying image quality. Therefore, this study aims to implement a deep learning model based on the U-Net architecture for monocular dataset segmentation and evaluate the model's performance under various image conditions presented. The objective of this research is to implement a deep learning model based on the U-Net architecture in monocular dataset segmentation and evaluate the model's performance against various image conditions such as changing lighting, the presence of shadows, and the diversity of objects and backgrounds in the image capture environment.

This research also aims to apply optimization techniques such as data augmentation, transfer learning, and training parameter adjustment to improve segmentation efficiency and accuracy. This approach has been proven to improve model performance in various domains, including medical image segmentation, aerial image segmentation, and microscopic image segmentation[6]. With this strategy, it is hoped that the U-Net model can perform accurate segmentation even under image conditions with high visual complexity and limited annotation.

II. RESEARCH METHODOLOGY



Img. 1 Research Process

Img. 1 shows the stages involved in the research. The method applied in this research is a segmentation-based deep learning method that focuses on using the U-Net architecture.

A. Problem Identification

Problem identification is carried out to determine the main issues to be addressed in this study. Accurate monocular dataset segmentation is a challenge in various applications, including navigation systems and image analysis. Conventional methods often have limitations in handling the complexity of road environments. Therefore, a deep learning-based approach is needed, particularly with the U-Net architecture, which has proven effective in various image segmentation tasks.

B. Literature Review

At this stage, a literature review was conducted on various previous studies related to image segmentation, deep learning, and the implementation of the U-Net architecture. This study aims to understand the technological developments and methods that have been used previously in monocular dataset segmentation in the context of roadways. The literature reviewed includes journals, scientific articles, and technical references related to deep learning and digital image processing. By understanding existing approaches, this research can determine the best strategy to achieve optimal results.

C. Needs Analysis

Needs analysis is necessary to determine the details required for research on segmentation using U-Net architecture on monocular datasets. This research uses a deep learning algorithm run on Visual Studio Code. Therefore, it requires a tool that can support the needs so that this research can run according to its objectives. The following are the requirements, divided into several parts:

- Hardware

  The hardware used in this study to achieve the research objectives is as follows:

  Processor :Intel(R) Core(TM) i5-12450H 2.00GHz
  VGA :NVIDIA GeForce RTX3050 4GB GDDR6
  RAM :16.00 GB
  System :Window 11 Home 64-bit
- Software

  The software used to operate the model in this study is as follows:

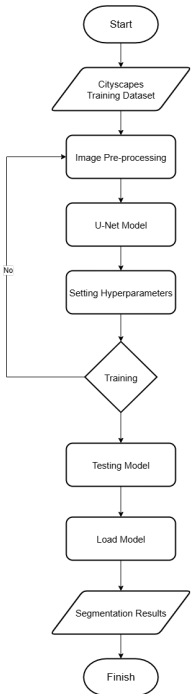
  • Google Chrome

  Designed for open source web browsers to obtain personalized search results.

  • Google Colab

  Used as a code editor medium to build the necessary architecture, Python is the language used in creating and designing programs.

D. System Design



Img. 2 U-Net Flowchart

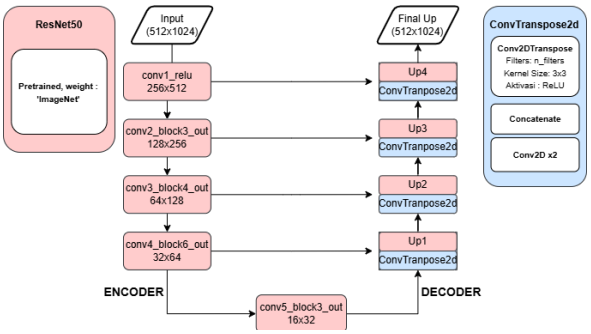
Img 2 shows the stages of designing monocular dataset segmentation using the U-Net architecture.

- *Cityscapes* Training Dataset

  The dataset was loaded and prepared as training data for the U-Net model.
- Image Pre-processing

  The images in the dataset are preprocessed, including resizing, normalization, and format conversion to match the U-Net model input.

- U-Net model



Img. 3 U-Net Model

The U-Net model in Fig. 3 was selected as the architecture for image segmentation. U-Net is a convolutional neural network (CNN) model that is widely used in segmentation tasks because it has an effective encoder-decoder structure with a ResNet50 backbone.

- Hyperparameter Settings
  The process of adjusting hyperparameters such as learning rate, batch size, and the number of epochs used to train the model.
- Training
  The model was trained using the processed Cityscapes dataset. If the model has not achieved the desired performance, this process can be repeated by adjusting the hyperparameters.
- Testing Model
  Model yang telah dilatih diuji menggunakan dataset evaluasi untuk mengevaluasi kinerja segmentasi dengan metrik.
- Load Model
  If the training process is successful and the model is good enough, the trained model is loaded for further testing.
- Segmentation Results
  The model is displayed and compared with the true mask to produce an image segmentation comparison showing the model's prediction of objects in the image.

E. System Implementation

The implementation was carried out using the following steps: Model training was performed using a processed dataset. The model was trained using appropriate optimization, such as the Adam optimizer, and adjusted hyperparameters, such as learning rate and number of epochs. During training, evaluation metrics such as mean Intersection over Union (mIoU) were used to measure model performance. If necessary, fine-tuning was performed to improve segmentation accuracy.

F. System Testing

P At this stage, the segmentation results are compared with the ground truth to measure the accuracy of the model. Testing is carried out in various scenarios to see the extent to which the model is able to adapt to varying road conditions. The results of this testing will be the main indicator in determining the success of the developed model.

G. System Evaluation

The evaluation was conducted by comparing the segmentation results with other methods, such as rule-based methods or other machine learning methods. If the results obtained were not optimal, improvements were made by fine-tuning the model. This evaluation also considered the computational efficiency of the model so that it could be applied in systems with limited resources. In addition, an error analysis was conducted to understand the error patterns that arose during the segmentation process.

III. *RESEARCH RESULTS*

This section presents the main findings obtained from the data analysis conducted in accordance with the methodology described earlier. The data was compiled and interpreted to answer the research questions and test the hypotheses proposed. The results are presented systematically through tables, graphs, and narratives to provide a clear and objective picture of the phenomenon under study.

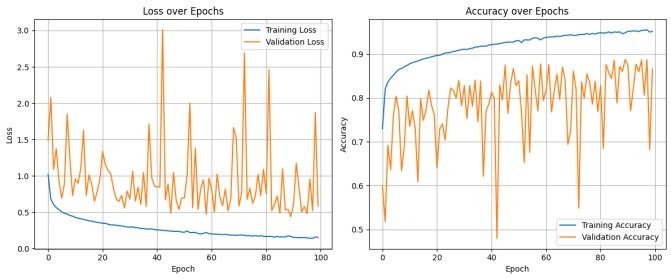
A. *Evaluation Results*

The experiment was conducted by testing five variations of learning rate values: 1e-3, 1e-4, 1e-5, 1e-6, and 1e-7, to see how much they affect the model's performance. The best results were obtained at a learning rate of 1e-4, with a Mean IoU value of 86.59%, pixel accuracy of 97.63%, and a final loss value of 0.0655. This indicates that at this value, the model can learn quickly to understand the data while remaining stable during the training process. Conversely, if the learning rate is too high, such as 1e-3, the model becomes unstable and the final results are poor, as it makes frequent large changes to the weights, making it difficult to find the correct patterns. On the other hand, at very low values such as 1e-6 and 1e-7, the updates made by the model are too small, making it difficult for the model to truly understand the data. Based on this testing, a learning rate of 1e-4 can be concluded as the best choice for training the U-Net model with ResNet50 on the Cityscapes dataset in the configuration used. The results of each test are presented in Table 1.

TABLE I
TEST RESULTS FOR EACH LEARNING RATE

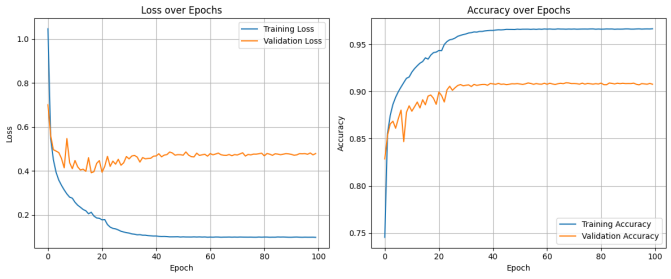
No	Learning Rate	Mean IoU (%)	Pixel Accuracy (%)	Final Loss
1	1e-3	60.32	89.22	0.1534
2	1e-4	86.59	97.63	0.0655
3	1e-5	77.94	96.23	0.1116
4	1e-6	32.76	88.63	0.3996
5	1e-7	14.04	76.22	1.0435

To strengthen the analysis of the influence of the learning rate, the following graph shows the training results of the U-Net model with the ResNet50 backbone on the Cityscapes dataset for each learning rate value, namely 1e-3, 1e-4, 1e-5, 1e-6, and 1e-7. Each graph consists of two parts: loss over epochs and accuracy over epochs, which show the changes in loss and accuracy values on both the training and validation data over 100 epochs. The explanations for each graph are as follows:



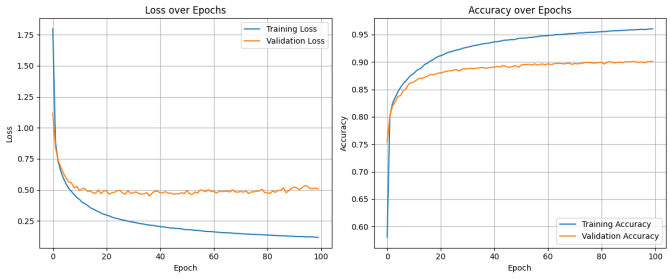
Img. 4 Learning Rate Graph 1e-3

The model with a learning rate of 1e-3 in Img. 4 shows unstable training. The validation loss is highly volatile and sometimes spikes sharply, even though the training loss continues to decrease. The validation accuracy also does not show consistent improvement, but rather fluctuates extremely. This indicates that the model has difficulty learning stably due to the learning rate being too large.



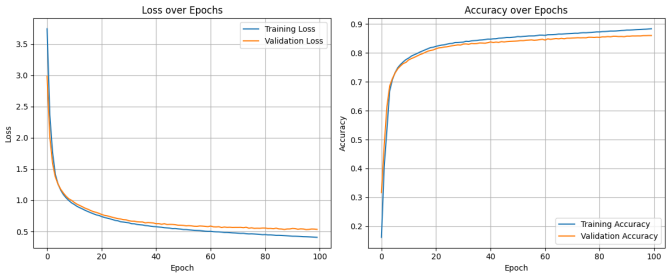
Img. 5 Learning Rate Graph 1e-4

The graph at a learning rate of 1e-4 in Img. 5 shows the best and most stable results. The training loss decreases consistently, and the validation loss is also relatively stable. The accuracy curves for the training and validation data increase over time and do not show significant signs of overfitting. This indicates that this value is the most optimal learning rate.



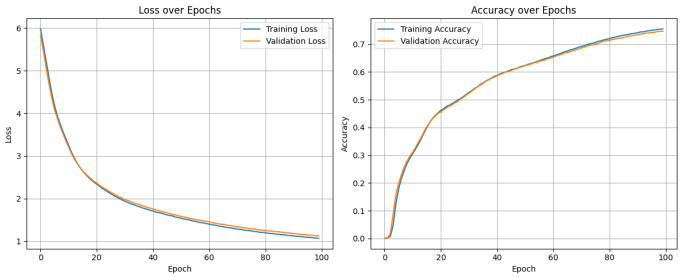
Img. 6 Learning Rate Graph 1e-5

With a learning rate of 1e-5 in Img. 6, the training process ran quite well, but the results were not as optimal as when using 1e-4. The training loss continued to decline consistently, indicating that the model was able to learn the training data well. However, on the validation data, the loss decline stopped more quickly and did not change much after several epochs. Additionally, a discrepancy begins to emerge between training and validation accuracy, indicating that the model is starting to overfit to the training data. This condition also suggests that the model's ability to recognize patterns in new data is not developing as quickly as when using a more appropriate learning rate.



Img. 7 Learning Rate Graph 1e-6

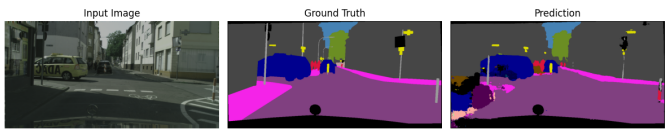
In the graph with a learning rate of 1e-6 in Img. 7, the model shows a very limited learning process. The decrease in training loss and increase in accuracy are very slow, and even appear to stop developing after the first few epochs. This indicates that a learning rate that is too small makes each weight update step very small, making it difficult for the model to adapt and understand the patterns in the data properly.



Img. 8 Learning Rate Graph 1e-7

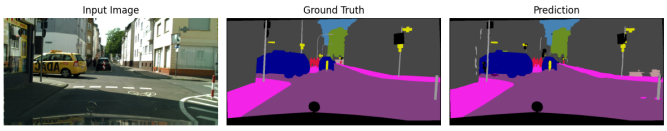
The model with a learning rate of 1e-7 in Img. 8 shows almost no significant progress during training. The loss and accuracy values only undergo slight changes and tend to remain at the same level throughout most of the epochs. This indicates that the learning rate used is too small, resulting in minimal weight updates that are insufficient to help the model effectively understand the data patterns. As a result, the training process does not proceed smoothly, and the model fails to achieve the expected performance.

B. *Visualization of Results*



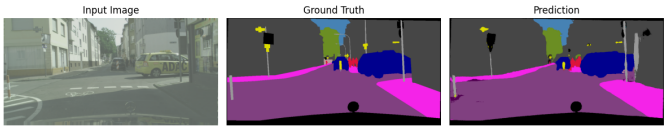
Img. 9 Visualization Results of Learning Rate 1e-3

In Img. 9, with a learning rate of 1e-3, the prediction results appear very unstable and tend to be chaotic. Some objects, such as vehicles, road markings, and traffic signs, are difficult to recognize accurately. Certain areas show coloring that does not match the intended class and display distracting visual artifacts. This indicates that the model is learning too aggressively, failing to form coherent segmentation, and lacking the ability to capture spatial features well.



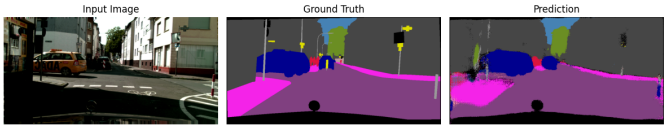
Img. 10 Visualization Results of Learning Rate 1e-4

The best segmentation results are shown in Img. 10 when the model is trained with a learning rate of 1e-4. In this configuration, almost all important elements in the image, such as roads, sidewalks, vehicles, and backgrounds, are successfully recognized and mapped accurately. Class colors are appropriate, boundaries between objects are clear, and prediction distributions are consistent with ground truth. This shows that the model is able to capture spatial representations optimally and generalize well to validation data.



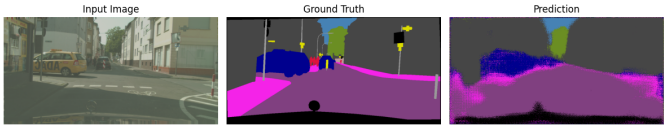
Img. 11 Visualization Results of Learning Rate 1e-5

In Img. 11, with a learning rate of 1e-5, the segmentation quality is still quite good but not as precise as the previous configuration. Some parts of objects such as poles or cars appear blurred or incomplete, and small details are sometimes overlooked. Although the segmentation is still visually acceptable, there is a decrease in terms of the accuracy of the shape and integrity of the objects.



Img. 12 Visualization Results of Learning Rate 1e-6

The use of a learning rate of 1e-6 in Img. 12 results in segmentation that tends to be less sharp. Several important areas, such as vehicles or road markings, are not predicted clearly, and the predicted colors appear to deviate from the labels. Although the model does not completely fail, visual details begin to disappear and its visual accuracy decreases dramatically.



Img. 13 Visualization Results of Learning Rate 1e-7

In Img. 13, with a learning rate of 1e-7, the model fails to produce meaningful segmentation. The prediction image appears as noise without recognizable object shapes or boundaries. Most of the area is randomly colored without following the visual pattern of the input.

This indicates that the model is barely learning at all because the weight updates are too small to capture relevant features during training. Overall, these visual results reinforce the numerical evaluation conclusion that a learning rate of 1e-4 is the most ideal configuration for producing accurate and stable segmentation on the Cityscapes dataset, while values that are too high or too low result in performance degradation in terms of both object shape and class accuracy.

IV. CONCLUSION

This study aims to evaluate the performance of the U-Net architecture with a ResNet50 backbone on semantic segmentation tasks using a high-resolution monocular dataset. The model was implemented on the Cityscapes dataset, which has good annotation quality and a large amount of data.

The experimental results show that U-Net with ResNet50 successfully produces accurate and stable object segmentation. With a learning rate configuration of 1e-4, the model achieves optimal performance with a Mean IoU of 86.59%, pixel accuracy of 97.63%, and final loss of 0.0655. These results demonstrate that the U-Net architecture is highly effective in learning complex spatial patterns in high-resolution image data, especially when supported by proper preprocessing, augmentation, and hyperparameter selection.

Factors such as image resolution, dataset size, annotation quality, and learning rate settings have been proven to significantly influence segmentation quality. This study emphasizes that the success of semantic segmentation depends not only on the model architecture but also on training strategies and the characteristics of the data used.

With the right approach, the ResNet50-based U-Net can be adapted for various real-world image segmentation needs, particularly in high-quality monocular data scenarios.

The following recommendations can be considered for further development:

1. Research Development
   It is recommended to explore other architectures such as DeepLabv3+, PSPNet, or transformer-based models, expand the training data with more diverse augmentations, and apply post-processing or ensemble methods to improve segmentation quality.
2. Real-world implementation
   The U-Net model can be applied to various needs such as road mapping and navigation systems, but it needs to be adapted to field data conditions and hardware limitations, including the use of pretrained models and efficient inference techniques.
3. Technical Aspects
   The selection of parameters such as learning rate greatly affects model performance. The use of learning rate schedulers, early stopping, and regularization techniques such as dropout and batch normalization can help prevent overfitting, especially on small

datasets with the support of transfer learning and additional augmentation.

REFERENCES

[1] S. K. Hussein and K. H. Ali, "Enhanced Semantic Segmentation of Aerial images with Spatial Smoothness Using CRF Model," *Al-Muthanna 2nd International Conference on Engineering Science and Technology, MICEST 2022 - Proceedings*, no. August, pp. 118–124, 2022, doi: 10.1109/MICEST54286.2022.9790187.

[2] A. Spolti *et al.*, "Application of U-net and auto-encoder to the road/non-road classification of aerial imagery in urban environments," *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, no. Visigrapp 2020, pp. 607–614, 2020, doi: 10.5220/0009337306070614.

[3] J. Straka and I. Gruber, "Modernized Training of U-Net for Aerial Semantic Segmentation," 2024.

[4] M. I. Ahmed, M. Foysal, M. Das Chaity, and A. B. M. A. Hossain, "DeepRoadNet : A deep residual based segmentation network for road map detection from remote aerial image," *IET Image Process*, vol. 18, no. September 2023, pp. 265–279, 2024, doi: 10.1049/ipr2.12948.

[5] M. Ciecholewski, "Urban scene semantic segmentation using the U-Net model," vol. 35, pp. 907–912, 2023, doi: 10.15439/2023F3686.

[6] D. Cheng and E. Y. Lam, "Transfer Learning U-Net Deep Learning for Lung Ultrasound Segmentation," 2021.