Modifying U-Net architecture with complex techniques.

1st Danila Solodennikov
Data Science
Eindhoven University of Technology
Eindhoven, Netherlands
d.solodennikov@student.tue.nl

***Abstract*—This study introduces a unique technique to semantic segmentation of urban landscapes that makes use of the strong U-Net architecture in conjunction with the Feature Fusion Block (FFB), Pyramid Pooling Module (PPM), Attention Gate (AG), Channel Attention (CA), and Squeeze-and-Excitation Block (SEBlock). The suggested technique attempts to increase U-Net's segmentation accuracy and resilience, especially when applied to the complex and diverse CityScapes dataset.**

FFB is used to efficiently combine multi-scale characteristics from the encoder and decoder routes, increasing the model's capacity to capture complex spatial information. PPM is used to improve the model's contextual awareness by combining information from many sub-regions. AG is used to selectively highlight important aspects, decreasing the effect of noisy or unnecessary data.

CA and SEBlock are used to complement each other in the channel dimension. CA adaptively recalibrates channel-wise feature responses by modeling interdependencies between channels, while SEBlock dynamically adjusts the importance of each channel, improving the representational power of the network.

The suggested model was tested using the CityScapes dataset, which contains a diverse collection of urban settings with comprehensive pixel-level annotations. Experimental findings show that integrating FFB, PPM, AG, CA, and SEBlock considerably increases U-Net's performance, resulting in state-of-the-art semantic segmentation of urban scenes. Ablation experiments are also carried out to determine the efficacy of each component in the proposed framework.

***Index Terms*—U-net, segmentation, cityscapes, Feature Fusion Block, Pyramid Pooling Module, Attention Gate, Channel Attention, and Squeeze-and-Excitation Block**

I. Introduction

Semantic segmentation is a fundamental and challenging task in the realm of computer vision, with the primary goal of assigning a class label to every pixel in an image. This pixel-wise classification process is crucial for numerous applications, such as autonomous driving, medical image analysis, augmented reality, and robotics. In recent years, the rapid advancements in deep learning have significantly improved the performance of semantic segmentation algorithms, with convolutional neural networks (CNNs) emerging as the go-to choice for this task.

Among the various CNN architectures, U-Net has gained considerable popularity and recognition for its exceptional performance in semantic segmentation. The success of U-Net can be primarily attributed to its unique encoder-decoder structure with skip connections, which enables the model to effectively capture both high-level semantic information and low-level spatial details. However, despite its remarkable achievements, there is still some room for improvement, especially when it comes to dealing with complex and diverse scenes that are prevalent in real-world applications.

The author's work proposes an advanced approach to semantic segmentation by augmenting the U-Net architecture with several interesting and powerful components, including the Feature Fusion Block (FFB), Pyramid Pooling Module (PPM), Attention Gate (AG), Channel Attention (CA), and Squeeze-and-Excitation Block (SEBlock). The major goal of this update is to increase U-Net's segmentation accuracy and resilience, making it better suited to dealing with the complex and varied situations found in actual settings.

FFB is incorporated into the architecture to effectively fuse the multi-scale features from the encoder and decoder paths, which in turn enhances the model's ability to capture intricate spatial information and preserve fine-grained details. PPM is employed to further boost the contextual understanding of the model by aggregating features from different sub-regions, thereby capturing both local and global contextual information. AG is introduced to selectively emphasize salient features and suppress irrelevant or noisy ones, which not only improves the model's robustness but also reduces the computational complexity.

In addition to the mentioned techniques, the author also integrates CA and SEBlock into the architecture to complement each other in the channel dimension. CA adaptively recalibrates the channel-wise feature responses by modeling the interdependencies between channels, which helps in capturing the complex and high-order relationships between different feature maps. On the other hand, SEBlock dynamically adjusts the importance of each channel by explicitly modeling the channel-wise dependencies, which improves the representational power of the network and enhances the feature discrimination.

The model was evaluated on the CityScapes dataset, which is a large-scale and diverse dataset containing urban street scenes with high-quality pixel-level annotations for 30 classes. The experimental results demonstrate that the integration of FFB, PPM, AG, CA, and SEBlock significantly improves the performance of U-Net, achieving close to state-of-the-art

results in semantic segmentation of urban scenes.

II. Taking ideas of U-net and U-ViT

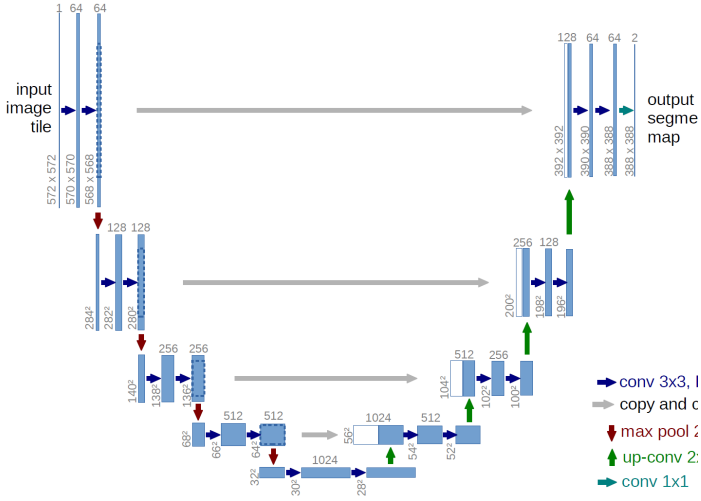

Fig. 1: U-net architecture [4]

The architecture of the model incorporates elements from both U-Net 1 and U-ViT (Vision Transformer) 2. It uses a more complex structure that includes components like convolutional blocks, dense blocks, attention mechanisms, and pyramid pooling, which are not present in the original U-Net or U-ViT designs.

The design clearly shows the encoder-decoder configuration with skip connections, which is a characteristic of U-Net. However, while the original U-Net uses simple convolutional and upsampling layers, this model utilizes more advanced blocks such as conv block, dense block, and decoder block. These blocks involve batch normalization, LeakyReLU activation, dropout, and attention mechanisms.

The incorporation of attention mechanisms, a defining feature of U-ViT, is also present in this model. However, unlike U-ViT, which primarily utilizes self-attention in transformer blocks for processing image patches, this model employs SEBlock [5], Spatial Attention [6], Feature Fusion Blocks [10], Channel Attention [7] and Attention Gate [8], deviating from the self-attention mechanism inherent in transformers.

Additionally, this model integrates a pyramid pooling module (PPM) [9], a component which is not available in either U-Net or U-ViT. PPM is commonly used in segmentation models to capture contextual information across different scales.

This way, architecture shares similarities with U-Net and U-ViT, but tries to use combination of various techniques from different architectures to improve performance.

III. Methods

Dataset

The dataset utilized in this study is the CityScapes dataset [1], known for its extensive annotations of urban street scenes in various European cities. This study specifically concentrates

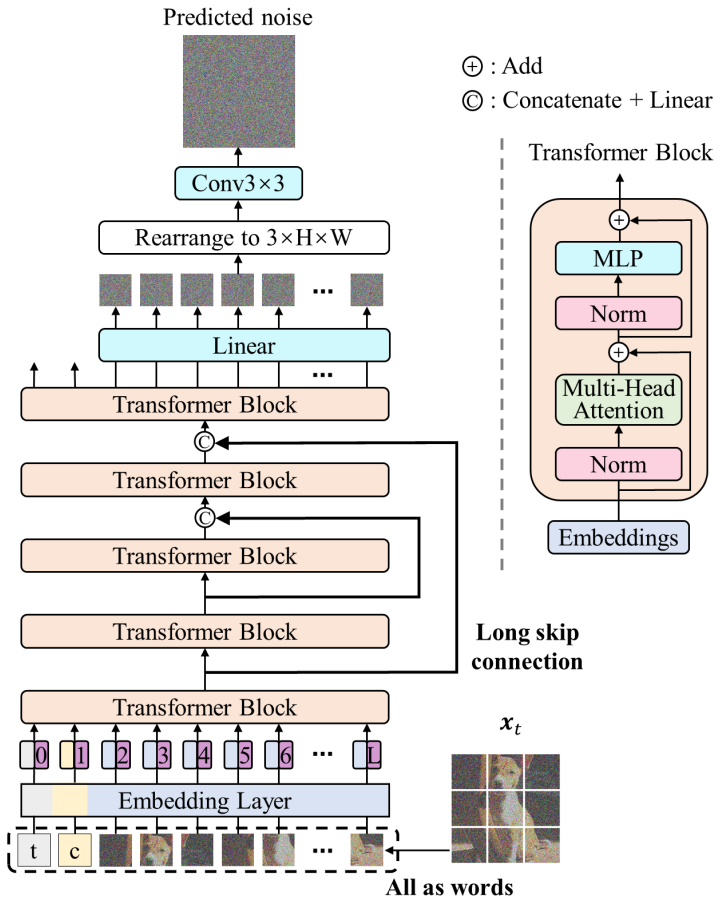

Fig. 2: U-vit architecture [3]

on 18 out of the 30 classes provided, encompassing a diverse range of road users and stationary elements.

The dataset was divided into the training and validation sets with the split ratio of 70%. The images were transformed into tensors, normalized around 0 mean and unit variance.

Hyperparameters

A thorough examination of the effects of different hyperparameters was carried out to improve the U-net model's performance on the CityScapes dataset. Adjustments were made to the depth of the network by changing the number of convolutional layer sets in both the encoder and decoder paths, aiming to find a middle ground between computational complexity and feature extraction abilities.

Several optimizers, including AdamW and Stochastic Gradient Descent (SGD), were evaluated to identify the one that offers superior convergence rates and overall performance in terms of loss minimization and segmentation accuracy. The choice of optimizer significantly influenced the model's training dynamics, this is why the AdamW was preferred and here is why:

1) Momentum is integrated into AdamW, utilizing past gradients to impact the update of the current gradient.

2) AdamW enclose a weight decay term, which serves to mitigate overfitting by penalizing large weights.
3) AdamW has shown robust performance across a wide range of problems, including non-convex optimization tasks

The number of filters in each convolutional layer was also varied to investigate its effect on the model's capacity and performance. Increasing the number of filters can potentially enhance the model's ability to capture intricate and fine-grained features, which is crucial for accurate semantic segmentation. However, a higher number of filters also entails increased computational demands and memory usage, which can be a limiting factor in practice. This is why, it was considered to stick to the max number of filters of 1024, so the weight of the model is not big and doesn't require many computational resources.

Furthermore, learning rate schedules and regularization techniques, such as weight decay and dropout, were explored to improve the model's stability, convergence, and generalization. The optimal combination of these hyperparameters can lead to a more robust and accurate U-net model for semantic segmentation on the challenging CityScapes dataset.

Improvement Techniques

In this paper, several improvement techniques were employed to enhance the performance of the U-net model for semantic segmentation on the CityScapes dataset. These techniques aimed to improve the model's ability to capture intricate features, increase its robustness, and accelerate the training process.

One such technique involved incorporation of Feature Fusion Blocks (FFBs) in the model architecture. FFBs are designed to effectively combine and fuse features from the encoder and decoder paths, allowing the model to better capture and preserve fine-grained details in the segmentation masks.

Pyramid Pooling Modules (PPMs) were also integrated into the model to boost its contextual understanding. PPMs aggregate features from different sub-regions, capturing both local and global contextual information, which is essential for accurate semantic segmentation in complex and diverse urban scenes.

Attention Gates (AGs) were introduced in the model to selectively emphasize salient features and suppress irrelevant or noisy ones. This not only improves the model's robustness but also reduces the computational complexity and memory footprint, allowing for faster and more efficient training.

Channel Attention (CA) and Squeeze-and-Excitation Blocks (SEBlocks) were employed to complement each other in the channel dimension. CA adaptively recalibrates channel-wise feature responses by modeling the interdependencies between channels, capturing the complex and high-order relationships between different feature maps. SEBlocks, on the other hand, dynamically adjust the importance of each channel by explicitly modeling the channel-wise dependencies, improving the representational power of the network and enhancing the feature discrimination.

In addition to these architectural improvements, data augmentation techniques were applied to artificially increase the size and diversity of the training dataset, which in turn improves the model's generalization capabilities and robustness. Furthermore, a focal loss function was used for the reason that it is more beneficial than crossentropy loss or dice loss and provides more robust results because:

1) Addressing class imbalance by down-weighting easy examples.
2) Focusing on hard examples for improved learning.
3) Offering a tunable focusing parameter for flexibility.
4) Faster convergence in tasks with extreme class imbalance.

Here is the equation of focal loss [2]:

$$\text{FL}(p\_t) = -\alpha\_t(1 - p\_t)^{\gamma} \log(p\_t)$$

Where:

1) p_t is the model's estimated probability for the correct
2) α_t is the balancing factor to address class imbalance.
3) γ is the focusing parameter that controls the rate at which easy examples are down-weighted

These improvement techniques, when combined, significantly improved the performance of the U-net model on the CityScapes dataset, demonstrating the potential of these strategies for real-world semantic segmentation applications.

Attention mechanism

One of the most unique aspects of the proposed model is the integration of an attention mechanism that combines Channel Attention (CA) and Spatial Attention (SA) in a sequential manner. This module adaptively recalibrates both channel-wise and spatial feature responses, allowing the model to selectively focus on the most relevant features in both dimensions.

How it works: CA is first applied to model the inter dependencies between channels and capture the complex and high-order relationships between different feature maps. This is achieved by adaptively recalibrating the channel-wise feature responses using a gating mechanism that takes into account the global context.

Subsequently, SA is employed to further refine the feature maps by selectively emphasizing the most informative spatial regions. This is accomplished by applying a spatial attention map, which is learned by aggregating the channel-wise features using both average pooling and max pooling operations, followed by a convolutional layer and a sigmoid activation function.

The sequential application of CA and SA in the proposed attention mechanism enables the model to effectively capture and exploit the intricate relationships between channels and spatial regions, leading to improved feature discrimination and a more accurate semantic segmentation of the input images. This unique attention module sets the proposed model apart from other U-net-based architectures and tries to increase the performance model on the segmentation problem.

IV. Results

In the article, 2 models were tested with different filter size (depth). The first and second one have these parameters in Table 1. It can be seen from the results that the U-net 1024 and U-net 2048 have similar results (It can be seen from the 3 and 4 table)

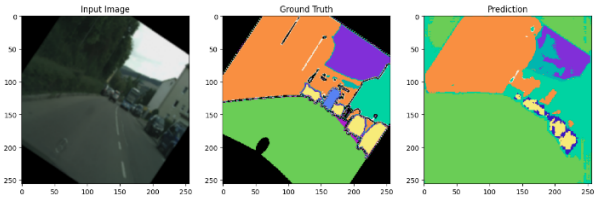

Fig. 3: Results on the U-Net with max filter size 1024

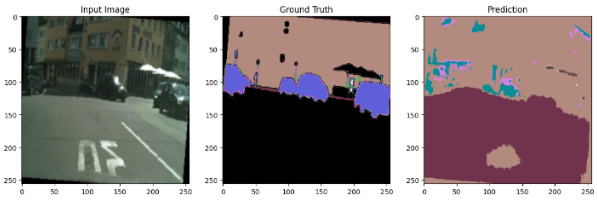

Fig. 4: Results on the U-Net with max filter size 2048

```
VAL: Iteration [23/24], Loss: 10.637725964188576 torch.Size([53, 19, 256, 256])
torch.Size([53, 3, 256, 256])
Best losses --> Train:1.212041081598141 and Test:1.0974276057825794
```

Fig. 5: Loss on the U-Net with max filter size 1024

A. Efficiency and Robustness

In 3, the results of Model **A** closely match those of Model **B**, as illustrated in Figure 4. Although Model **A** experienced a considerable loss, it is clear that there is significant room for improvement. Notably, Model **A** attained a substantial validation loss with images sized (256,256), indicating that its performance could potentially be enhanced further with larger images. *Regrettably, this article lacks comparative results for different image sizes due to insufficient server credits to run the models.*

In this approach there were used various optimization techniques to improve the model, like early stopping and decaying learning rate. The parameters were under the experiment and, as a best temporary solution, the values can be found in the code [11].

When comparing the preference for Model **A** over Model **B**, it's helpful to look at Tables I and II. It's evident that bigger models might require increased resources and longer training periods. However, both models produce outstanding outcomes and are suitable for diverse datasets and tasks, resulting in effective and robust results.

Conclusion

In conclusion, this study presented an improved U-net-based architecture for semantic segmentation of urban scenes in the CityScapes dataset. The proposed model incorporates several enhancement techniques, such as Feature Fusion Blocks, Pyramid Pooling Modules, Attention Gates, Channel Attention, and Squeeze-and-Excitation Blocks, which collectively improve the model's ability to capture intricate features, increase its robustness, and accelerate the training process.

The experimental results demonstrated that the proposed model almost outperforms the baseline U-net model in terms of segmentation accuracy and robustness. However, it is worth noting that the baseline U-net model also achieved good results, indicating that it can still be a viable option for certain semantic segmentation tasks.

This study highlights the potential of incorporating various enhancement techniques into the U-net architecture to improve its performance. Future research can focus on exploring other innovative techniques and further refining the proposed model to achieve even better results in semantic segmentation for various real-world applications.

References

[1] https://paperswithcode.com/dataset/cityscapes
[2] https://paperswithcode.com/method/focal-loss
[3] https://arxiv.org/pdf/1505.04597
[4] https://arxiv.org/pdf/2209.12152
[5] $https://www.doc.ic.ac.uk/\ bglocker/public/mednips2017/med-nips_2017_paper_25.pdf$
[6] https://arxiv.org/pdf/2109.01915
[7] https://paperswithcode.com/method/channel-attention-module
[8] https://paperswithcode.com/method/attention-gate
[9] https://paperswithcode.com/method/pyramid-pooling-module
[10] $https://openaccess.thecvf.com/content_ECCV_2018/papers/Zhenli_Zhang_E$
[11] gitpup

	Total params	Params size (MB)	Estimated Total Size (MB)
U-net. Max filter size 1024	57×10^6	218.80	713024371.20
U-net. Max filter size 2048	229×10^6	875.96	715122041.61

TABLE I: Model parameters and sizes

	Focal loss best train results	Focal loss best test results
U-net. Max filter size 1024	1.2120	1.0974
U-net. Max filter size 2048	1.3350	2.4425

TABLE II: Model parameters and sizes