Dean&Francis

ISSN 2959-6157

Investigating the Impact of Residual Connections and the Integration of VGG19 Architecture on U-Net for Car View Segmentation

Yunyang Wang

School of Mathematics and Statistics, University of St. Andrews, St Andrews, United Kingdom

*Corresponding author: Yw219@st-andrews.ac.uk

Abstract:

This study investigates the performance of three computer vision neural networks architecture, which are the standard U-Net, Deep Residual U-Net(ResU-Net), and VGG19 Integrated U-Net (VGG19U-Net) on car view segmentation. The models are trained with 4000 images and their masks and are tested at different stages of training. The validation criterion includes training and validation loss, Intersection of unions and Dice coefficient. The results demonstrate that ResU-Net outperforms the other models in segmentation accuracy while maintaining competitive prediction speeds. The VGG19U-Net shows improved performance over the standard U-Net, highlighting the benefits of deeper architectures in semantic segmentation tasks. Additionally, the research underlines the importance of architectural modifications like residual connections and deeper convolutional layers for enhancing segmentation accuracy. This study offers valuable insights into optimizing U-Net variants for vehicle segmentation, which can be extended to other real-world applications, including autonomous driving. These findings provide a view for future improvement in real-time image segmentation for complex environments.

Keywords: Deep learning; Sematic segmentation; Computer vision; U-Net.

1. Introduction

Autonomous driving and advanced driver assistance systems (ADAS) has become the crucial functions for vehicles [1]. And image segmentation, car view segmentation in particular, is one of its core techniques for the system to capture the road and surrounding information.

Unlike more traditional object detection or image classification, which only need to capture and classify a given set of targets or classes, the image segmentation aims to segment every distinct object

in the image [2]. While it can provide far more abundant and detailed information, the computational complexity become higher. Also, as the segmentation task is used for self-driving, the prediction accuracy and speed are critical. Traditional bounding box regression network, like Faster R-CNN or You Only Look Once (YOLO) are no longer feasible for this task, and more complex network, like Mask CNN or Fully Convolutional Network (FCN) replace their job [3]. The more advanced U-Net architecture, originally developed for biomedical image segmentation, has proven to be effective across various domains [4]. And there is a growing need to explore modifications and enhancements to the standard U-Net architecture.

This study aims to investigate the impact of two significant modifications to the U-Net architecture for car view segmentation:

(1) Integrate the VGG19 architecture, known for its depth and strong feature extraction capabilities [5].

(2) Corporate the residual connections, which have shown promise in addressing the vanishing gradient problem in deep networks [6].

By training and comparing the standard U-Net, VGG19U-Net, and ResU-Net, this research seeks to understand how these architectural changes affect segmentation performance, training dynamics, and prediction speed [7, 8]. The study utilizes a dataset of 5000 front view images collected by self-driving cars, providing a realistic and challenging benchmark for evaluating these models, giving feasible solutions for self-driving.

2. Methodology

The purpose of this paper is to evaluate the feasibility and advantageous of the standard U-Net, ResU-Net and VGG19U-Net, from their speed of prediction and multiple criterions for accuracy.

For the evaluation, the dataset is sourced from Kaggle, which contains 1000 front view images collected by self-driving cars and segmented. The image and their masks are paired and transformed for uniformity and readiness. Also, regularization is applied to prevent from gradient vanishing and increase stability [7].

2.1 U-Net

The standard U-Net is a convolutional neural network modified and extended from the FCN [3]. Instead of the single CNN layer to build the mask in FCN, it replaces it with an expansive convolutional network for up sampling, which makes it more feasible for semantic segmentation [4]. The network down samples with two 3x3 convolutional layers followed by an activation function, and a max pooling for feature extraction. At each step the feature maps are cropped and copied to combine with each corresponding step in up sampling after the 2x2 up-conv [4].

2.2 VGG19U-Net

VGG19 is a 19 layers convolutional neural network for classification task, containing 5 blocks in the backbone with 2 convolutional layers in each of the first 2 blocks, and 4 convolutional layers in the last 3 blocks [5]. The three fully connected layer in the head turns the feature map into vector for classification [5].

VGG19 adds 3 more layers from VGG16 and performs better with higher accuracy and precision, indicating that adding convolutional layers could increasing the network performance for certain dataset [9].

From this assumption, as the standard U-Net only contains 2 convolutional layers in each down sampling step, VGG19 is integrated by replacing the four encoder blocks in original U-Net to the five blocks in VGG19. As shown in Fig. 1, the first four blocks are still cropped and concatenate with the corresponding up sampling block. For symmetry, the fifth block of the backbone serves as bottleneck to pass the feature from encoder to decoder.

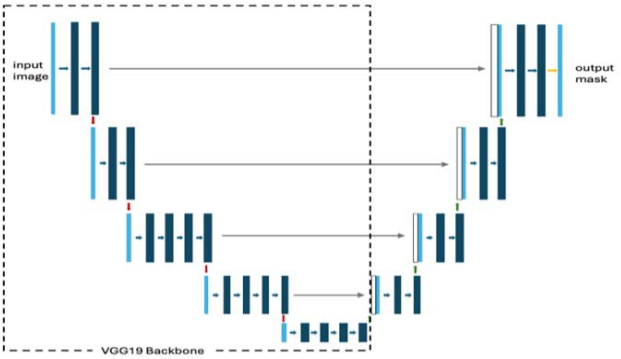

Fig. 1 Architecture of U-Net with VGG19 Backbone

ISSN 2959-6157

2.3 ResU-Net

Deep Residual Network, or ResNet, is a neural network in strong capability for deep learning. The residual connection between convolutional layers allows the gradient to flow back during back propagation, which effectively avoiding the vanishing gradient problem when number of layers is large [6]. Also, the Skip Connections allow the model to decide whether to update, compensating for the irreversible information loss caused by high nonlinearity [6].

ResUNet that accumulates the advantage of both U-Net and ResNet is introduced for investigation. For each block in both encoder and decoder, a residual connection is added. Also, like the standard U-Net, the ResUNet uses skip connections to concatenate the feature maps from the encoder with those in the decoder. Unlike the ResUNet in Road Extraction by Deep Residual U-Net, no blocks are deleted, to control the experimental variables [8].

3. Training

Hyper parameters are tunned during the training to maximize the model performance. No overfitting are met during training so no weight decay is used. Loss plateaus are often met during training during epochs 15 to 25, especially in the VGG19UNet. This problem often occurs in the way what the loss remained unchanged for a few epochs in the plateau and explode to a high value and not decreasing even after multiple epochs of training, so ReduceLROnPlateau is used radically with patience=1 and factor=0.5, where after one epoch with loss non-decreasing, the learning rate immediately decrease by multiple the factor 0.5 for punishment. Adam algorithm is used as optimizer, which is computed as,

$$m_t ? \leftarrow ? \beta_1 ? m_{t-1} ? + ? (1 ? - ? \beta_1) ? g_t \quad (1)$$

$$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (2)$$

$$\widehat{m_t} \leftarrow m_t / (1 - \beta_1^t) \quad (3)$$

$$\widehat{v_t} \leftarrow v_t / (1 - \beta_2^t) \quad (4)$$

$$\theta_t \leftarrow \theta_{t-1} - \gamma \widehat{m_t} / \left(\sqrt{\widehat{v_t}} + ? \right) \quad (5)$$

where β1,β2 are double momentum(β1,β2) to update the biased mean(*mt*) and variance(*mt*), and moving unbiased means(*mt*) and variances(*vt*) of past gradients(θt−1) to current gradients(θ_t), to guide future updates more effectively[10].

4. Results

The performance of the three models - U-Net, VGG19U-Net, and ResU-Net - was evaluated using multiple criteria, including training and validation loss, Dice coefficient, Intersection over Union (IoU), and prediction speed.

4.1 Training and Validation Loss

Fig. 2 The train loss(left) and validation loss(right) for the three models to train 40 epochs

The training and validation loss curves for U-Net, VGG19U-Net, and ResU-Net are shown in Fig. 2, which compares model across 40 epochs.

All three models show rapid convergence during the first 5 epochs at different rates. The standard U-Net begins with a relatively higher loss (starting at around 1.2) and converges more slowly compared to the other models. VGG19U-Net and ResU-Net has a low initial loss and quickly reducing to below 0.1 within the first 10 epochs. ResU-Net, in particular, continues to maintain a slightly lower loss throughout training comparing to another two models, indicating that the residual connections help stabilize the learning process by vanishing gradients, especially in deeper layers[11].

The validation loss trend mirrors that of training loss, where three models show a generally decreasing trend.

ResU-Net exhibits the lowest validation loss, with minimal fluctuations, stabilizing near 0.05. On the other hand, VGG19UNet performs comparably with similar trend to ResU-Net early on but shows a noticeable spike around epoch 25, where its validation loss rises before decreasing again. This spike could indicate a temporary overfitting issue or the model attempting to adjust to more complex features as training progresses caused by the more complex and deeper VGG19 structure. And the strike is fixed immediately by the optimizer.

4.2 Dice Coefficient and IoU

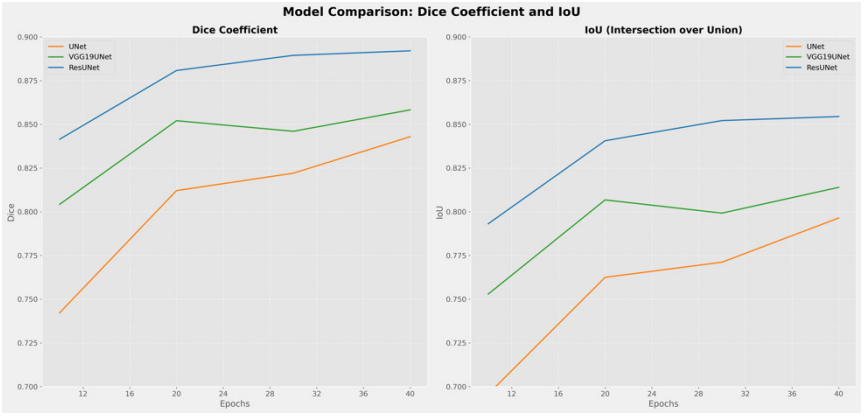

Fig. 3 Model Comparison: Dice Coefficient(left) and IoU of the three models training for epochs 10, 20, 30 and 40

Dice coefficient and IoU are two different criterions to measure the pixel accuracy in Fig. 3. Dice coefficient measure the similarity between two sets (predicted and ground truth) with formula $(2\cdot|A\cap B|)/(|A|+|B|)$. It ranges from 0 to 1, where 1 indicates perfect overlap. IoU also computes the overlap between predicted and ground truth masks, but with formula $|A\cap B|/|A\cup B|$.

Both measurement shows similar increasing trend in general. ResU-Net outperforms both U-Net and VGG19UNet, achieving a relatively high accuracy even in only the first 10 epochs of training, and reach a Dice Coefficient close to 0.9, which demonstrates the effectiveness of residual connections in capturing more nuanced spatial features of the car segmentation task. However, even the ResU-Net is able to capture features rapidly, the model improvement becomes gradual, so further tunning is needed to continue increase accuracy. VGG19UNet achieves a final Dice Coefficient of around 0.85, indicating better performance than the standard U-Net. VGG19UNet's deeper architecture enables it to better extract multi-scale features, which contribute to more accurate segmentation. The drop of the accuracy line in both measurement matches the strike in the loss plot, which suggests that the network's complexity may have introduced some instability. This could be addressed by further adjustment of hyperparameters, like introducing weight decay. The unmodified U-Net performed the worst among the models, but even between 30 and 40 epochs, the model's accuracy continued to improve significantly. This suggests that, under the current parameters, there is still substantial room for learning, even though its learning ability is not as fast as the other two models.

Table 1. Average time spend for each model to predict the mask for one image

	U-Net	VGG19U-Net	ResU-Net
Average prediction time per image	0.0041(s)	0.0043(s)	0.0055(s)

4.3 Prediction Speed

As shown in Table 1 In prediction speed, U-Net is the fastest, with an average prediction time of 0.0041 seconds per image. VGG19UNet follows closely with an average time of 0.0043 seconds, while ResU-Net is slightly slower at 0.0055 seconds per image. While the added complexity of residual connections in ResU-Net, or more layers and convolutional structure in VGG19U-Net slightly increases the inference time, this is compensated by its improved segmentation accuracy.

And Fig. 4 illustrates examples of the target mask and the

ISSN 2959-6157

mask predicted by three models at 40 epochs of training.

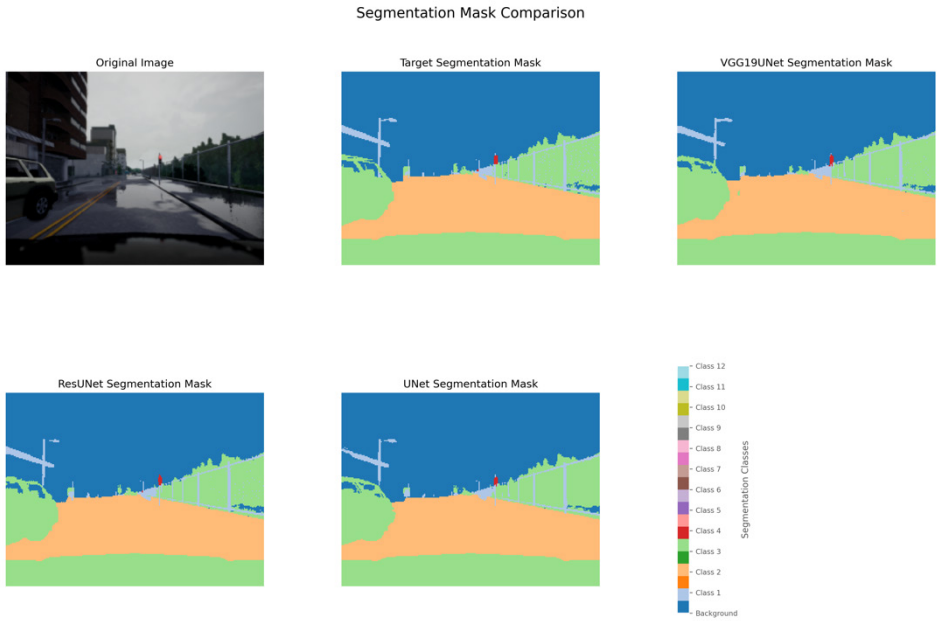

Fig. 4 Illustration of the result for the three models training at 40 epochs, with the original image and target mask

5. Conclusion

This study has comparatively analysed the U-Net variants for car view segmentation and yields several significant findings. The ResU-Net architecture demonstrated superior performance, achieving the highest Dice coefficient and IoU scores, and a comparable prediction speed. The most rapid convergence rate and most stable learning curve indicates the advantage of residual connections in feature capturing and learning. Also, The VGG19U-Net showed significant improvement over the standard U-Net with close training speed at the same time, underlining the benefits of more convolutional and pooling layers for capturing features in segmentation tasks. However, the occasional training strike in VGG19U-Net causes the accuracy failing to achieve the expectations, suggest potential for further optimization. However, the standard U-Net, while performing less impressively, shows consistent improvement even in later training stages, indicating untapped potentials with extended training.

These results make implications for the field of autonomous driving and ADAS. The improved accuracy brought by ResU-Net and VGG19U-Net enables a deeper understanding of the road and better object recognition in autonomous driving applications, which significantly boost the safety of self-driving vehicles particularly in complex urban environments where precise segmentation is needed. Also, the low trade-off between the precision and prediction speed suggest that these two networks have great potentials in real time prediction for self-driving.

However, the prediction speed, even for the fastest U-Net, of 0.0041s/image(frame) is still not sufficient for real time tasks. 0.246 second is needed for one second of a 60fps video, and the previous image processing, and afterwards localizing and mapping, path planning and decision making are as well time-consuming tasks. The delay from information capture to decision make will lead to fail of response to risk. So further optimization in network architecture and convolutional block design are needed. Also, more diverse datasets would provide insights into their generalization capabilities in various driving conditions.

In conclusion, this study not only advances the understanding of U-Net variants in car view segmentation but also contributes to the broader goal of enhancing autonomous vehicle technologies. As the automotive industry continues its rapid evolution towards autonomy, research in this domain will play an important role in shaping the future of transportation, promising safer roads and more efficient urban mobility.

References

[1] Van Brummelen J, O'Brien M, Gruyer D, et al. Autonomous vehicle perception: The technology of today and tomorrow. Transportation Research Part C: Emerging Technologies, 2018, 89: 384-406. https://doi.org/10.1016/j.trc.2018.02.012

[2] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
[3] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651. https://doi.org/10.1109/tpami.2016.2572683
[4] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science, 2015: 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Computer Vision and Pattern Recognition, 2014. http://export.arxiv.org/pdf/1409.1556
[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. CVPR, 2016. https://doi.org/10.1109/cvpr.2016.90
[7] Nawaz A, Akram U, Salam AA, et al. VGG-UNET for brain tumor segmentation and ensemble model for survival prediction. 2021 International Conference on Robotics and Automation in Industry (ICRAI), 2021: 1-6. https://doi.org/10.1109/ICRAI54018.2021.9651367
[8] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753. https://doi.org/10.1109/lgrs.2018.2802944
[9] Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for image classification. 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), 2021: 96-99.
[10] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv.org, 2014. https://doi.org/10.48550/arXiv.1412.6980
[11] Balduzzi D, Frean M, Leary L, et al. The shattered gradients problem: If resnets are the answer, then what is the question? International Conference on Machine Learning, 2017: 342-350.