Project 5 - Amodal Part Segmentation

Zhenjie Jiang
zhejiang@student.ethz.ch

Shuaijun Gao
shugao@student.ethz.ch

Xintian Yuan
xinyuan@student.ethz.ch

ABSTRACT

In Amodal Part Segmentation, we predict the segmentation of different parts of the hand including the parts that are not visible from the image. After we experimented with various models, losses and data augmentations, we used an altered U-Net model with ResNet 50 as backbone, Dice-CE loss and modified data augmentation. We were able to achieve a score of 0.5617 on the public leader board.
Keywords: U-Net, ResNet 50, Dice-CE loss

1 INTRODUCTION

In traditional semantic segmentation tasks, only the visible parts of the objects are predicted, however human can easily identify partially occluded object, and predict its true shape. Thus, for computer vision models to imitate more of human capabilities we would like to also predict the occluded sections of the objects. Such ability is becoming increasingly important in the real world with the advancement of technology, for example in autonomous driving, the ability to recognize occluded objects is essential for autonomous vehicles to detect dangers and make good decisions.

In this paper, we focus on recognising different parts of the left and right hand in hand images. In our experiments, we used fully-supervised methods, and focused on tuning three parts of the experiments: model architecture, loss function and data augmentation. We adapted existing segmentation models U-Net and ResNet into a new model, and used combined dice loss and cross-entropy loss functions for amodal semantic segmentation and apply data augmentation to achieve strong performance on the given dataset.

2 RELATED WORK

U-Net[11] U-Net is network capable of capturing context through contracting path, and accurate localization through a symmetric expanding path. The training strategy used in the paper relied on strong data augmentation, and the cross-entropy energy function which is calculated pixel-wise. Thanks to the data-argumentation with elastic deformation, U-Net was able to perform very well with only very few annotated images, and demonstrated large improvements over previous models. While U-Net achieved great segmentation performances, it required prior knowledge regarding the depth of the model and only allowed the restrictive fusion of feature maps at the same scale[16]. U-Net is limited in extracting image-derived information that is related to the location of an object with a non-standard shape.

Semantic Amodal Segmentation[17] This paper introduced the task of Semantic Amodal Segmentation to encourage machine vision systems to gain human-vision capabilities. They annotated two datasets, one smaller dataset with 500 images from the BSDS dataset to compare annotation statistics, and one larger dataset with 5000 images from COCO dataset for developing algorithms. The paper discovered that the task is well defined, and annotations are consistent across annotators. They used IoU to compare the regions and F scores to measure the edges. The paper also developed two new metrics for this tasks, which are amodal segment quality and pairwise depth ordering between regions. They also provided strong baselines according to their metrics.

A survey of loss functions for semantic segmentation[3] This paper evaluated some of the popular loss functions for Image Segmentation, as well as introduced a new loss function called log-cosh loss. The paper describes the formulas of the popular loss functions such as Binary Cross-Entropy, Dice Loss, Focal Loss, Combo Loss (Dice + CE) etc. The paper also stated their use cases, for example, Binary Cross-Entropy works best for balanced data, Dice Loss works best for imbalanced data. The paper conducted experiments with a 2D U-Net model and discovered that no loss functions had the best performance in all tests, and it depended on the data-set for the majority of the cases.

UNet++: A Nested U-Net Architecture[16] UNet++ re-designed skip connections between the encoder and decoder in order to reduce the semantic gap between the feature maps of the encoder and decoder sub-networks.They inserted dense convolution block whose number of convolution layers depends on the pyramid level before the feature maps of the encoder being received by the decoder. They also used deep supervision[5] method which enables more accurate segmentation for parts that appear at multiple scales.

3 METHOD

In our experiments, we split the dataset into training set, and validation set. Each experiment is trained on the training set for 24 hours, and evaluated on the validation set. The scores on the leaderboard are substantially higher than the scores in the experiments, since for the leaderboard we trained the model on the entire dataset instead of just the training set for the experiments.

Initially we analysed the performance of the provided skeleton code, which used U-Net and cross-entropy loss. The model achieved a score of 0.375 on the validation set. Although the model was able to capture the general shape of the hands, the model struggled to differentiate each finger clearly. The model also failed to distinguish the hands when two hands are overlapped.

3.1 Model

U-Net's inability to capture the shape of each finger coincides with the weakness stated in the original U-Net paper [11], that it struggled with non-standard shapes in images. Thus we increased the complexity of the U-Net to better capture the shape of the fingers. After adding one extra encoder layer and one extra decoder layer, the model was able to separate the hands much better. However, the model was still unable to properly distinguish the fingers, and it sometimes identifies hands that did not exist in the image.

Next, we focused on improving the decoder side only, we decided to add attention mechanism into the decoder layer [9]. We hypothesised that the attention mechanism should help the model

to focus on the finger structures. Although not perfect, the model improved the segmentation on the fingers, and it alleviated the issue of identifying non-existing hands.

Finally, we arrived at using ResNet [2] as the backbone (encoder) for U-Net, as we think ResNet should provide better ability in extracting smaller features such as different parts of the finger. For the decoder part of the network, after reviewing many different approaches, we decided to use the method where each decoder layer predicts a section of the final result, and concatenate the outputs of each decoder layer as the final output, which should retain the ability of focusing on the finer structure of the image from the attention mechanism. The model showed great improvement over previous models in all areas, which in particular ResNet 50 demonstrated the best performance as the backbone.

3.2 Loss

For the loss function, we tested various loss functions and discovered that dice plus cross-entropy loss performed the best. Dice Loss [6] is essentially a measure of overlapping between the prediction and the ground truth, dice loss dynamically assign fewer weights to avoid strong influences from the easy samples, which allows better prediction for harder images such as the images where there are a lot of occluded parts. The additional cross-entropy loss helped in increasing the stability of the loss, as it allows more diversity in the loss function.

3.3 Data Augmentation

For data augmentation, we tested many different sets of parameters on U-Net with ResNet 50 backbone. We discovered that by setting transition and scale factor over 0.3, there is a significant improvement in the miou score. Although increasing the color factor did not have a noticeable effect on the score, we still decided to keep the factor over 0.3 to make it in line with the other factors. We found that decreasing the rotation factor had a negative impact on the score, which is likely due to the level of occlusion is heavily affected by the rotation of the hands. So We kept the rotation factor at the default value of 5.

In the end, we settled with the values of 0.3, 0.3, 5, 0.3 for transition, scale, rotation, and color factors respectively.

4 EVALUATION

We conducted the experiments according to the steps defined in section 3.

First, we fixed the loss functions and data augmentation hyper-parameters, and then tried with different model architectures. After choosing the best network, we changed the combination of loss functions and found the one with highest score. In the last step, we experimented on different data augmentation hyper-parameters.

As shown in the Figure 3, the different variants of U-Net achieved the best results among all the other models. And the top variants are U-Net with ResNet as the backbone (named U-Net AE), which used ResNet as the encoder for the original U-Net.

U-Net with ResNet 50 achieved 0.4419129021990573 on the validation data set and 0.561721599099 on the public evaluation set. The corresponding loss function was dice loss plus cross-entropy. The data augmentation hyper-parameters were 0.3, 0.3, 5, 0.3 for transition, scale, rotation, and color factors respectively. The whole training process lasted for 34 hours on the Leonhard cluster.

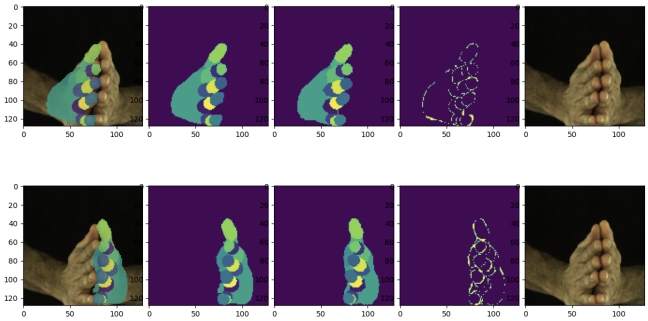

Figure 1: One intermediate prediction of hand amodal part segmentation.

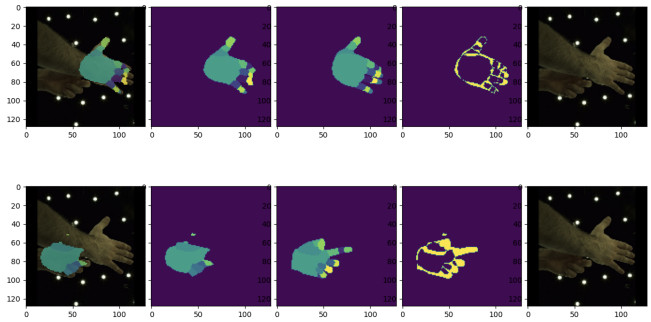

Figure 2: One intermediate prediction of more complex models.

While U-Net with ResNet 50 performed quite well and had a promising up trend on the validation dataset and the public leaderboard, by carefully assessing the intermediate pictures, we were able to identify several potential future improvements for this model.

As we can see in the Figure 1, although the model was capable of predicting the main body part of the two hands quite accurately, the prediction on the edges of the hand were still lacking, which shows that the model were still lacking in terms of modelling unusual shapes.

5 DISCUSSION

As is mentioned in section 3, we experimented with a lot of different model architectures, loss functions, and data augmentation parameters. However, not all of them improved the segmentation performance. We would like to discuss in depth about the potential causes of the underwhelming results of these failed experiments.

We started with trying several more complex segmentation models, such as U-Net++[16], Deeplab[1], Denseaspp[13], PSPnet[15], and Refinet[7], because we thought that increasing the model complexity will help improving the encoding of the details of the hands. But none of them actually surpassed the performance of the original U-Net model. After looking into the real predictions of these

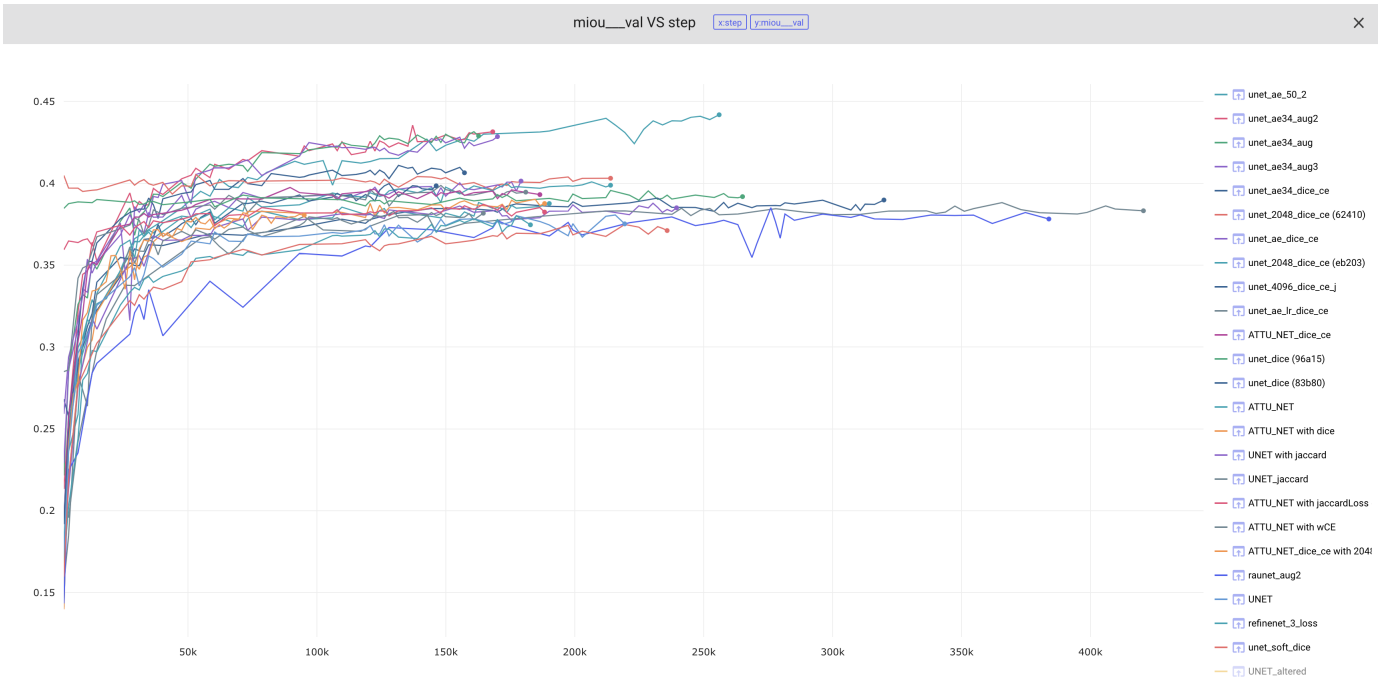

Figure 3: All experiments with different model architectures, loss functions and data augmentation hyperparameters.

models, we finally found out that they lacked accurate predictions of invisible parts of the hands.

As shown in the Figure 2, the unseen part of the hand in the bottom was omitted by the model. The lack of generalization of these complex models gave rise to the tendency of over-fitting to the visible parts.

We also experimented with various loss functions and combinations of loss functions, namely dice loss[12], focal loss[8], cross entropy[14], jaccard loss[10] and dice with top k loss[4]. Although dice loss and cross entropy achieve great performance, other loss functions don't. Focal loss adds a discount factor to the standard cross entropy criterion, but it degenerates to normal cross entropy with balanced foreground-background labels. Jaccard loss can not estimate the importance of the hidden part of the hands, which leads the model to only focus on visible parts.

6 CONCLUSION

We discovered that by changing the encoder of U-Net to ResNet 50 results in significant performance gain in both the prediction of the visible and the invisible parts of the hands, we think this was due to the increased ability in discovering smaller features, which the original encoder struggled. Furthermore, combining dice loss and cross entropy allowed the model to emphasize on occluded parts of the hand. Together with reasonable data augmentation techniques, we finally arrived at the completed model, which performed well on the hand amodal part segmentation task.

REFERENCES

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv:cs.CV/1606.00915

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:cs.CV/1512.03385

[3] Shruti Jadon. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 1–7. https://doi.org/10.1109/CIBCB48159.2020.9277638

[4] Maksim Lapin, Matthias Hein, and Bernt Schiele. 2016. Loss Functions for Top-k Error: Analysis and Insights. arXiv:stat.ML/1512.00486

[5] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. 2014. Deeply-Supervised Nets. arXiv:stat.ML/1409.5185

[6] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice Loss for Data-imbalanced NLP Tasks. arXiv:cs.CL/1911.02855

[7] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2016. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. arXiv:cs.CV/1611.06612

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal Loss for Dense Object Detection. arXiv:cs.CV/1708.02002

[9] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2018. Attention U-Net: Learning Where to Look for the Pancreas. arXiv:cs.CV/1804.03999

[10] Md Rahman and Yang Wang. 2016. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation, Vol. 10072. 234–244. https://doi.org/10.1007/978-3-319-50835-1_22

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:cs.CV/1505.04597

[12] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. 2017. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *Lecture Notes in Computer Science* (2017), 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

[13] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. 2018. DenseASPP for Semantic Segmentation in Street Scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3684–3692. https://doi.org/10.1109/CVPR.2018.00388

[14] Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. arXiv:cs.LG/1805.07836

[15] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. arXiv:cs.CV/1612.01105

[16] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. Unet++: Redesigning skip connections to exploit multiscale features

in image segmentation. *IEEE transactions on medical imaging* 39, 6 (2019), 1856–1867.

[17] Yan Zhu, Yuandong Tian, Dimitris Mexatas, and Piotr Dollár. 2016. Semantic Amodal Segmentation. arXiv:cs.CV/1509.01329