Selective Image Enhancement via Semantic Segmentation with Skin Tone Preservation

Jiaqi Sun
Department of Statistics
University of Michigan
sunjiaqi@umich.edu

***Abstract*—This project explores semantic segmentation for selective image editing, inspired by Apple's *Photographic Styles*, which preserves natural skin tones while enhancing other regions. A MobileViT+DeepLabV3 model was fine-tuned using the Human Parsing dataset. The model achieved a high overall accuracy of 94% and demonstrated superior segmentation of skin regions compared to the pre-trained model. The segmentation results were applied to selectively enhance image aesthetics using LUT filters, preserving skin tones while improving the background. This work highlights the potential of semantic segmentation for practical applications in image editing.**

***Index Terms*—Semantic segmentation, Image processing, Fine-tuning, Human parsing**

I. Introduction

Semantic segmentation, a foundational task in computer vision, involves partitioning an image into semantically meaningful regions at the pixel level. It has wide applications in fields like medical imaging and autonomous driving. The development of semantic segmentation can be divided into three major stages. Early methods relied on hand-crafted features such as edge detection [1] or color segmentation [2] [3]. While straightforward, these approaches struggled with complex images and image variations.

Recent advancements in this field was primarily driven by deep learning. The introduction of the Fully Convolutional Network (FCN) [4] marked a significant shift in 2015, as it enabled end-to-end segmentation by replacing fully connected layers with convolutional layers and leveraging upsampling techniques. U-Net [5] refined this approach by an encoder-decoder structure with ”short-cuts”, combining low-level feature maps (rich in detail) with high-level ones (rich in semantic information). This architecture improved the precision of segmentation masks and excelled in medical imaging tasks. Meanwhile, the DeepLab series [6] [7] [8] introduced multi-scale feature fusion and dilated convolutions, which enhanced the model's ability to adapt to complex and extended the receptive field without increasing parameters.

More recently, Transformer-based models have achieved state-of-the-art performance. Vision Transformer (ViT) [9] introduced self-attention mechanisms to capture global dependencies within an image. It introduced a tokenized approach to redefine image tasks in a manner similar to textual tasks. SegFormer [10] incorporated multi-level Transformer decoder blocks with multi-layer perceptrons (MLPs) to balance accuracy and efficiency. Swin Transformer [11] introduced a hierarchical sliding-window attention mechanism, which significantly reduced computational complexity. These advancements have positioned Transformer-based architectures as the dominant paradigm in semantic segmentation, achieving superior performance across various benchmarks.

This project aims to leverage semantic segmentation techniques to address a specific challenge in image editing. Apple introduced a feature called *Photographic Styles*, which allows users to adjust image colors selectively without altering skin tones. It relies on advanced hardware for real-time image segmentation when taking photos, making it unavailable on older iPhones. The aim is to replicate this functionality on existing images by fine-tuning a MobileViT+DeepLabV3 model on a Human Parsing dataset. The pipeline includes semantic segmentation to identify skin regions and selective application of Look-Up Table (LUT) filters to enhance non-skin areas.

II. Method

A. Dataset & Model

Semantic segmentation is a supervised learning problem. The input consists of high-resolution three-channel color images, and the output is a pixel-level segmentation mask of the same resolution, where each pixel's value corresponds to its category. A pre-trained model was fine-tuned to reduce the training burden, and the Great Lakes HPC at University of Michigan is utilized to accelerate the training process.

The Human Parsing dataset was used for fine-tuning. It contains 17,706 samples of fashionably dressed individuals. Each sample includes an original image (approximately 600×400 resolution) and a corresponding mask image. The masks are annotated with 18 categories that segment the pixels into regions such as background, hair, face, arms, and legs. This dataset is well-suited for the project, as its detailed human skin labels are essential for applying filters selectively to different image regions.

For the model, Apple's MobileViT+DeepLabV3-small was selected as the initial model. MobileViT, pre-trained on the PASCAL VOC dataset at a resolution of 512×512, is enhanced with a DeepLab V3 head for segmentation. This model strikes a balance between efficiency and accuracy, making it a strong foundation for the project. With a parameter size of only 6.4 million, it is lightweight and highly suitable for deployment on mobile devices with limited computational resources.

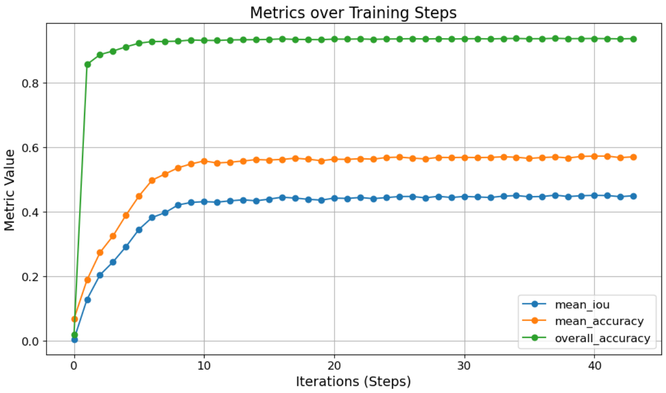


Fig. 1: Model performance on validation set during the fine-tuning process

B. Preprocessing, Training & Evaluation

To simulate real-world variations and improve model generalization, we implemented a series of data augmentation techniques using ‘albumentations‘ library. These augmentations included random horizontal flips, Gaussian noise, and image downscaling to simulate low-quality inputs. Adjustments to brightness, contrast and saturation were made to handle diverse lighting conditions, while structural distortions such as pixel dropout and were introduced to reflect real-world imperfections. These transformations were applied randomly to the training data, while the validation data underwent minimal preprocessing to ensure consistent evaluation.

Fine-tuning was conducted using the AdamW optimizer with an exponentially decaying learning rate scheduler. A weighted cross-entropy loss function was used, assigning higher weights to non-background categories. The model was trained for 10 epochs with a batch size of 8. During validation, metric mean_IoU was imported to evaluate the alignment between predicted and ground-truth masks.

After training, the fine-tuned model was saved and tested on unseen images to generate segmentation masks. To visually inspect the model's performance, the masks were overlayed on the input images. For aesthetic adjustment, a Look-Up Table (LUT) transformation filter was applied to non-skin regions, ensuring that skin tones remained natural while enhancing the background.

III. Results

A. Fine-Tuning Results

The metrics used to evaluate the performance of the fine-tuned model include mean Intersection over Union (mean IoU), mean accuracy, and overall accuracy.

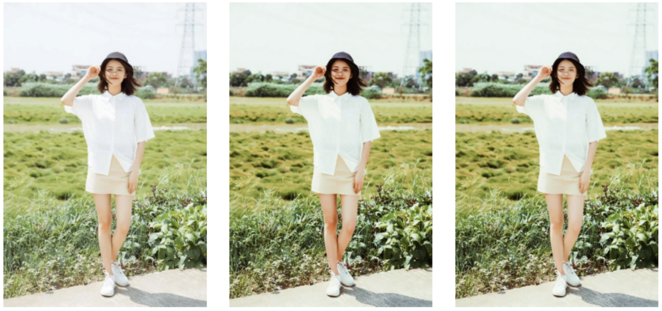
- Mean IoU: The average Intersection over Union ratio across all categories.
- Mean accuracy: The average proportion of correctly predicted pixels for each category.
- Overall accuracy: The proportion of correctly predicted pixels over the entire image.

Figure 1 illustrates the evolution of these metrics during the fine-tuning process. The overall accuracy stabilizes around 0.93-0.94, reflecting the model's strong and consistent performance across the entire dataset. However, mean IoU and mean accuracy exhibit relatively lower values due to the imbalanced pixel distribution across categories. This imbalance makes it challenging for the model to achieve high performance uniformly across all categories. The training and validation loss throughout the training process are shown in Appendix 1.

B. Inference & Application

After fine-tuning, the model exhibits significantly improved performance compared to the original pre-trained model. The original model was only capable of coarse segmentation, distinguishing between the person and the background in the image. In contrast, the fine-tuned model successfully segments detailed components of the human body, including the face, arms, and legs. The comparison images are shown in Appendix 2.

To further validate the application of semantic segmentation, a Kodak Portra 400 LUT filter was applied to enhance an example image (Figure 2). When the LUT was applied globally, the background colors became vivid and exposure was controlled. However, this adjustment negatively impacted the person's skin tone, rendering it dull and unnatural. After combining the segmentation results with the LUT application, the resulting image demonstrated a substantial improvement: the final image preserved natural skin tones while enhancing the background. This approach highlights the utility of semantic segmentation in creating aesthetically pleasing edits.



(a) Original image (b) Global filter (c) Filter on parts

Fig. 2: LUT filter on an example image

IV. Conclusion

This project developed a semantic segmentation pipeline for selective image adjustments, preserving skin tones while enhancing other regions. By fine-tuning MobileViT+DeepLabV3 on the Human Parsing dataset, the model achieved high accuracy and detailed segmentation of body parts. The selective application of LUT filters demonstrated the practical benefits of combining segmentation with targeted image enhancements, offering a robust solution for real-world applications in image editing and aesthetics.

The code for fine-tuning and applying LUTs can be found on GitHub.

REFERENCES

[1] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[2] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[3] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 282–289.

[4] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *CoRR*, vol. abs/1412.7062, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1996665

[7] L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *CoRR*, vol. abs/2105.15203, 2021. [Online]. Available: https://arxiv.org/abs/2105.15203

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

V. APPENDIX

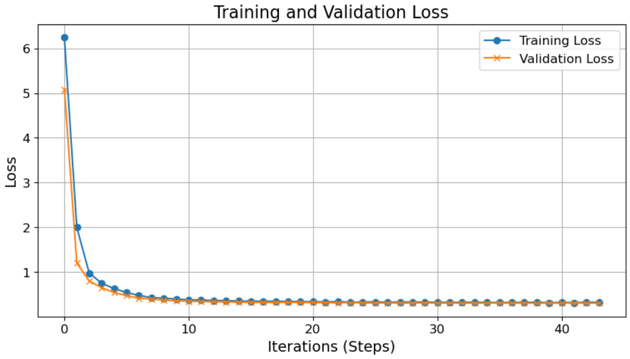
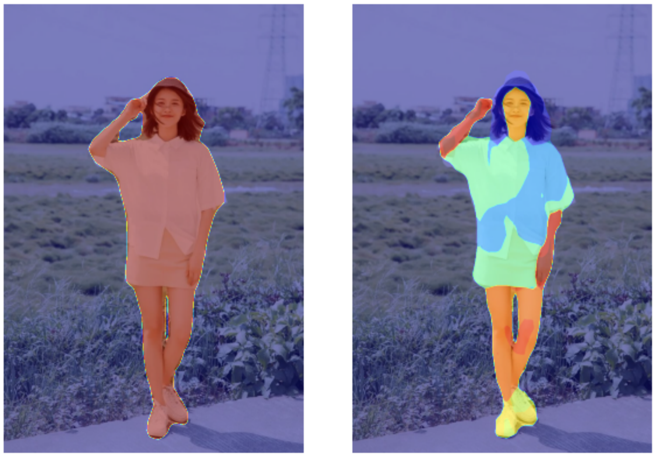


Fig. A1: Loss throughout the training process



(a) Original Model

(b) Fine-tuned model

Fig. A2: Performance comparison between original model and fine-tuned model