Report for CE7454 Project 1 CelebAMask Face Parsing

Zhang Saining
G2403905K

Abstract

*This report details the methodology for Project 1: CelebAMask Face Parsing, part of the CE7454 course. I present **LISA**, the Lightweight SkipNeXt for Face Parsing, which consists of only 1,999,551 parameters. The model is primarily based on the SegNeXt framework, incorporating the skip connection to connect the features from the shallow layer and the deep layer. Additionally, various data augmentation strategies and loss functions have been employed to improve overall performance. Ultimately, the proposed method achieves a mean Intersection over Union (mIoU) score of 57.84 on the CelebAMask Face Parsing test set.*

1. Introduction

This project is based on the well-known CelebAMask-HQ Dataset [4]. I would like to congratulate the CelebA team [5] for winning the PAMI Mark Everingham Prize at ECCV 2024 and express gratitude for their significant contributions to the open-source community. This section will provide a brief introduction to the dataset, outlining the project's requirements and the analysis of the project.

1.1. CelebAMask-HQ

CelebAMask-HQ (Fig.1) is a large-scale face image dataset that has 30,000 high-resolution face images selected from the CelebA dataset by following CelebA-HQ. Each image has segmentation mask of facial attributes corresponding to CelebA. The masks of CelebAMask-HQ were manually-annotated with the size of 512 x 512 and 19 classes including all facial components and accessories such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth.

CelebAMask-HQ can be used to train and evaluate algorithms of face parsing, face recognition, reconstruction, pose transfer, conditional image generation, face editing, and 3D-aware image synthesis.

1.2. Project Requirements

Face parsing assigns pixel-wise labels for each semantic components, e.g., eyes, nose, mouth. The goal of this mini challenge is to design and train a face parsing network. We will use the data from the CelebAMask-HQ Dataset. For this challenge, we prepared a mini-dataset, which consists of 5,000 training and 1,000 validation pairs of images, where both images and annotations have a resolution of 512 x 512.

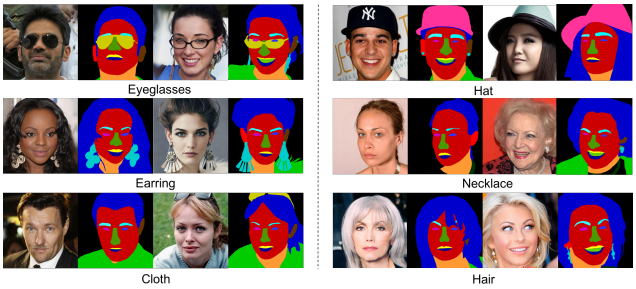

Figure 1. Sample images in CelebAMask-HQ.

The performance of the network will be evaluated based on the mIoU between the predicted masks and the ground truth of the test set (100 images). Here are some regulations:

- Train your network using our provided training set.
- Tune the hyper-parameters using our provided validation set.
- To maintain fairness, your model should contain fewer than 2,000,000 trainable parameters.
- No external data and pretrained models are allowed in this mini challenge. You are only allowed to train your models from scratch using the 5000 image pairs in our given training dataset.
- You should not use an ensemble of models.

1.3. Analysis

Based on the project requirements, it is evident that only 5,000 images are available for training, with no allowance for additional data, pretrained models, or knowledge distillation techniques. Therefore, implementing comprehensive data augmentation strategies is crucial to enhance the training dataset and develop a model with improved generalization capabilities.

Additionally, the model is constrained to a parameter limit of 2,000,000, highlighting the necessity of utilizing

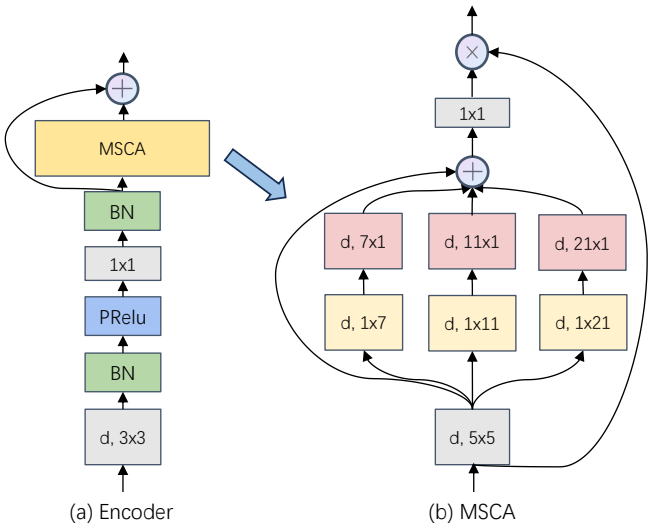

Figure 2. The structure of the encoder block and MSCA of LISA. 'd' stands for depthwise convolution.

Stage	Block	Ouput Size
e1	1	512×512×32
e2	2	256×256×64
e3	2	128×128×142
e4	3	64×64×320
e5	1	32×32×512
d4	1	64×64×320
d3	1	128×128×142
d2	1	256×256×64
d1	1	512×512×32
C	-	512×512×n

Table 1. The overall architecture of LISA. 'e' is the encoder, 'd' is the decoder, 'C' is the final 1×1 size convolution layer. 'n' is the number of classes, in this task, 'n' is 19.

lightweight networks or modules. Furthermore, evaluating various loss functions is also critical for optimizing model performance.

2. Methodology

This section provides a detailed overview of the methodology employed in this study. Sec.2.1 outlines the data augmentation techniques utilized, while Sec.2.2 describes the design and architecture of the network. Additionally, Sec.2.3 discusses the loss function implemented in the proposed method.

2.1. Data Augmentation

Given the limited dataset of 5,000 training images, effective data augmentation is crucial. Initially, I applied horizontal flipping to all images and corresponding labels, producing an additional 5,000 mirrored samples. It is essential to note that during this flipping process, the labels for the left and right eyes, as well as those for the eyebrows and ears, were appropriately swapped. Subsequently, I implemented random rotations (±10 degrees) and color jitter on the existing 10,000 images, resulting in an additional 20,000 images. This augmentation expanded the training set to a total of 30,000 training images. The effectiveness of these data augmentation strategies was further quantitatively assessed in Sec.3.

2.2. Network Structure

Due to the constraint of the model parameters, I employed the lightweight network for segmentation, SegNeXt [2], which uses multi-scale attention, as the baseline for further optimization. However, since SegNeXt concatenate outputs from multi stages together and go throught linear blocks and the Hamburger module [1] cause large parameter occupancy, it's hard to cut the number of parameters to 2 million while keep good results. Therefore, I combine SegNeXt with skip connection together to achieve cheaper computational cost and better performance. The proposed network is referred to as the Lightweight SkipNeXt for Face Parsing (LISA).

The architecture of LISA is shown in Tab.1. The pipeline is composed of five encoding stages, four decoding stages, and a single linear classification layer. During the encoding stage, I utilize multiple building blocks. As demonstrated in [2], the model can achieve effective results with a simplified decoder; therefore, I implement a single block in the decoding stage. Notably, a 2x downsampling operation is performed between each interval of encoding stages.

Fig.2 illustrates the structure of the encoder. I retain the multi-scale convolutional attention (MSCA) module from [2] to effectively fuse multi-scale convolutional features.

As shown in Fig.2 (b), the MSCA consists of three components: a depthwise convolution to aggregate local information, multi-branch depthwise strip convolutions to capture multi-scale context, and an 1×1 convolution to model relationships between different channels.

The output of the 1×1 convolution is utilized as attention weights to reweight the input of the MSCA.

After preliminary experiments, I discard some linear layers that had a limited impact on the model's performance and redesigned the entire encoder layer to light the model and keep the performance. Specifically, from Fig.2 (a), the encoder is composed by one 3×3 depthwise convolution block, one 1×1 pointwise convolution block and the

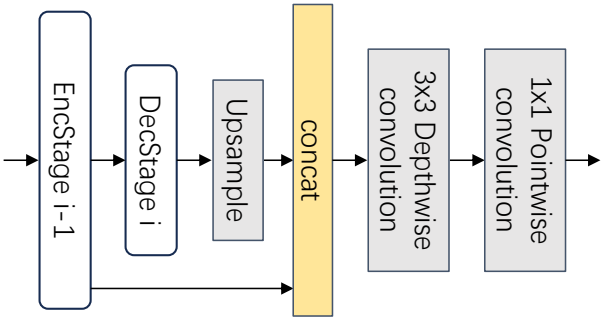

Figure 3. The structure of the decoder. When $i = 5$, the upsampling input is come from the Hamburger module.

MSCA block. In the encoder, I use PReLU [3] as the activation function to improve model fitting with nearly zero extra computational cost and little overfitting risk. This will also be discussed in Sec.3.

For decoder depicted in Fig.3, I first upsample the feature map from the stage i decoder and concatenate the upsampled feature map with the features from the stage $i-1$ encoder. The concatenated feature is then processed through depthwise and pointwise blocks, similar to the encoder, resulting in a feature map with the same number of channels as the output of the stage $i-1$ encoder. When $i = 5$, the input for upsampling is derived from a Hamburger module whose input is the stage 5 encoder's output. With this skip connection, it's allow us to fuse detailed features from the shallow layer with semantic features from the deep layer.

2.3. Loss Function

In this task, I employ the Cross Entropy Loss and the Dice Loss for optimization the parameters.

Cross Entropy Loss is widely used for multi-class classification problems, it is written like:

$$\mathcal{L}_{\text{CE}} = -\sum_{i} y_i \cdot \log(\hat{y}_i), \tag{1}$$

where y is the ground-truth label, and $\hat{y}$ is the prediction.

Dice Loss can effectively handle situations in segmentation tasks where there is a significant disparity in the number of positive and negative samples, making the model pay more attention to the segmentation effects of the minority classes:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y| + \lambda}. \tag{2}$$

where X is the ground-truth label, and Y is the prediction, λ is 0.0001 in the task.

	Val		Test	
	mIoU	F1 score	mIoU	Params
UNet	68.03	0.81	53.05	1,999,011
SegNeXt	65.65	0.83	51.53	1,992,819
LISA	75.04	0.88	57.84	1,999,551

Table 2. The comparison across UNet, SegNeXt and LISA with parameters under 2 million.

Flip	Rot	CJ	Val		Test
			mIoU	F1 score	mIoU
✗	✗	✗	71.64	0.84	56.08
✓	✗	✗	73.02	0.87	56.81
✓	✓	✗	74.16	0.88	57.16
✓	✓	✓	75.04	0.88	57.84

Table 3. The results for ablation studies of the data augmentation. 'Flip' is horizontal flip. 'Rot' is random rotation. 'CJ' is color jitter.

	Val		Test
	mIoU	F1 score	mIoU
Dice	71.33	0.84	55.89
CE	73.74	0.85	56.41
CE + Dice	75.04	0.88	57.84

Table 4. The results for ablation studies of the loss function. 'Dice' is Dice loss. 'CE' is Cross Entropy loss.

	Val		Test
	mIoU	F1 score	mIoU
ReLU	74.08	0.87	57.16
GeLU	74.55	0.88	57.45
PReLU	75.04	0.88	57.84

Table 5. The results for the comparison between different activation funtions.

3. Experiments

3.1. Implementations

In accordance with the project requirements, I trained SegNeXt and LISA models, each with fewer than 2 million parameters, for comparative analysis, alongside a UNet [6] utilizing skip connections. The training dataset comprised 30,000 images post data augmentation. All experiments were performed on an NVIDIA A800-SXM4-80GB, utilizing a batch size of 40 over 100 epochs, with early stopping implemented if no improvement in the best model was observed over 20 epochs. All experiments were evaluated using mIoU and F1 scores on the validation set.

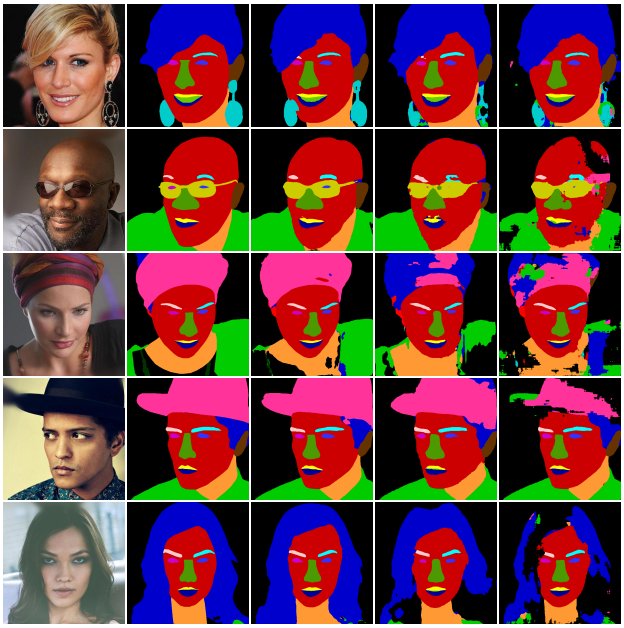

Figure 4. Qualitative results of different methods on the validation set. Left to right: images; ground-truth masks; LISA; SegNeXt; UNet.

3.2. Main Results

As illustrated in Tab.2, under identical parameter constraints and experimental conditions, the LISA model demonstrates a substantial improvement over the baseline SegNeXt and UNet. Specifically, on the validation set, LISA exceeds SegNeXt by 9.39 and 0.05 in mIoU and F1 score, respectively, and surpasses UNet by 7.01 and 0.07 in mIoU and F1 score. On the test set, LISA achieves 57.84 of mIoU, outperforming SegNeXt and UNet by 6.31 and 4.79, respectively.

From qualitative results in Fig.4, it is evident that the proposed LISA method significantly outperforms the baseline methods in both local categories (e.g., earrings, lip) and larger area categories (e.g., hat, hair, and cloth). These results clearly illustrate the superiority of LISA across various categories.

Fig.5 presents results obtained from the test set using LISA.

3.3. Ablation Studies

Efficacy of the data augmentation. From Tab.3, as the number of data augmentation methods increases, the metrics of LISA on both the validation and test sets are continuously improving. Compared to the method without data augmentation, the results trained with all data augmentation techniques are 3.40 (mIoU) and 0.04 (F1 score) higher on the validation set, and 1.76 (mIoU) higher on the test set.

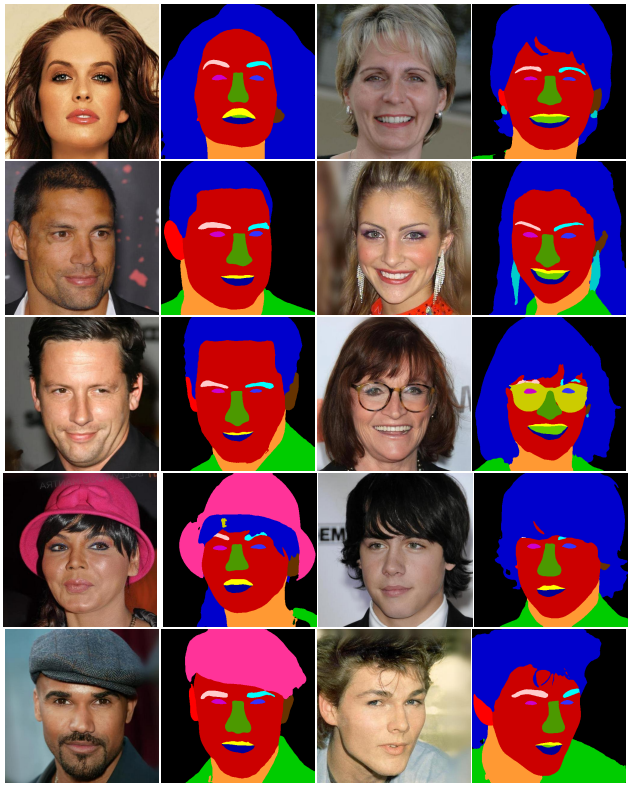

Figure 5. Some results on test set set using LISA.

These results indicate that the data augmentation methods effectively enhance the model's generalization ability.

Efficacy of the loss function. From Tab.4, when using the combination of Dice loss and Cross Entropy loss results in improved model performance compared to using either loss function in isolation. This suggests that, for this task, the two loss functions work synergistically to guide the model toward better convergence.

Efficacy of the activation function. From Tab.5, it can be observed that employing PReLU as the activation function yields slightly better results than using GeLU or traditional ReLU in the SegNeXt model. This improvement may be attributed to PReLU's ability to enhance the model's fitting capability without significantly increasing the risk of overfitting, thereby contributing to improved generalization performance.

4. Conclusion

In this report, I propose a method for Project 1: CelebAMask Face Parsing of the CE7454 course **LISA**. It extracts essence from the SegNeXt and add skip connection to connect the features from the shallow layer and the deep layer. Testing on the CelebAMask Face Parsing test set, LISA achieves 57.84 in mIoU.

Acknowledgement

All parts of the method is mainly designed by myself. I have discussed SegNeXt and skip connection with my friend in the lab, Li Peishuo.

References

[1] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? In *ICLR*, 2021. 2

[2] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3

[4] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3