

SegFace: Face Segmentation of Long-Tail Classes

Kartik Narayan, Vibashan VS, Vishal M. Patel

Johns Hopkins University
{knaraya4, vvishnu2, vpatel36}@jhu.edu

Abstract

Face parsing refers to the semantic segmentation of human faces into key facial regions such as eyes, nose, hair, etc. It serves as a prerequisite for various advanced applications, including face editing, face swapping, and facial makeup, which often require segmentation masks for classes like eyeglasses, hats, earrings, and necklaces. These infrequently occurring classes are called long-tail classes, which are overshadowed by more frequently occurring classes known as head classes. Existing methods, primarily CNN-based, tend to be dominated by head classes during training, resulting in suboptimal representation for long-tail classes. Previous works have largely overlooked the problem of poor segmentation performance of long-tail classes. To address this issue, we propose *SegFace*, a simple and efficient approach that uses a lightweight transformer-based model which utilizes learnable class-specific tokens. The transformer decoder leverages class-specific tokens, allowing each token to focus on its corresponding class, thereby enabling independent modeling of each class. The proposed approach improves the performance of long-tail classes, thereby boosting overall performance. To the best of our knowledge, *SegFace* is the first work to employ transformer models for face parsing. Moreover, our approach can be adapted for low-compute edge devices, achieving 95.96 FPS. We conduct extensive experiments demonstrating that *SegFace* significantly outperforms previous state-of-the-art models, achieving a mean F1 score of 88.96 (+2.82) on the CelebAMask-HQ dataset and 93.03 (+0.65) on the LaPa dataset.

Code — <https://github.com/Kartik-3004/SegFace>

1 Introduction

Face parsing, a semantic segmentation task, involves assigning pixel-level labels to a face image to distinguish key facial regions, such as the eyes, nose, hair, and ears. The identification of different facial regions is crucial for a variety of applications, including face swapping (Xu et al. 2022), face editing (Lee et al. 2020a), face generation (Zhang, Rao, and Agrawala 2023), face completion (Li et al. 2017), and facial makeup (Wan et al. 2022). Long-tail classes are those that occur infrequently within a dataset. Existing face parsing datasets (Lee et al. 2020a) consist of these long-tail

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

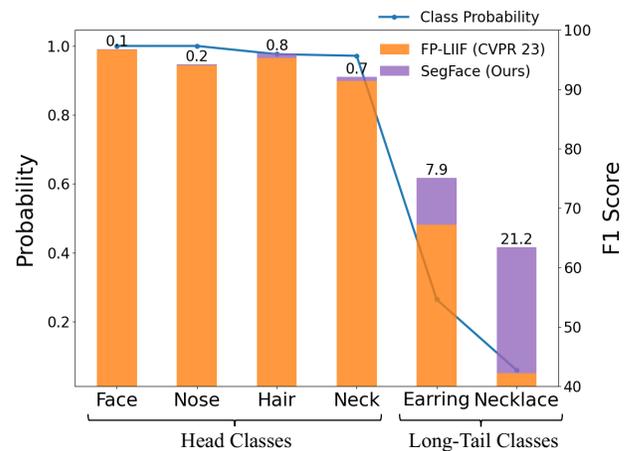


Figure 1: The proposed *SegFace* leverages a lightweight transformer decoder with learnable class-specific tokens. The association of each class with a token enables the independent modeling of each class, which boosts the segmentation performance of long-tail classes that typically underperform in existing methods. The blue line represents the probability of a class being present in a randomly selected image from the CelebAMask-HQ train set. *SegFace* provides a significant boost in the segmentation performance of long-tail classes (+7.9, +21.2), thereby establishing a new state-of-the-art in face parsing performance.

classes, which are mostly accessories like eyeglasses, necklaces, hats, and earrings, because not all faces will feature these items. We cannot expect to have equal representation of all classes in current or even future face-parsing datasets, as certain facial attributes like hair, nose and eyes are naturally more common than accessories like earrings and necklaces. Additionally, it is difficult to collect samples with less frequently occurring classes. Moreover, detailed annotation for face segmentation, especially for less common or smaller facial features, is labor-intensive and costly.

Since the advent of deep learning in semantic segmentation (Long, Shelhamer, and Darrell 2015), numerous studies have focused on solving face segmentation. Several works (Guo et al. 2018; Zhou, Hu, and Zhang 2015; Lin

et al. 2021) leverage the learning potential of deep convolutional neural networks to achieve promising face segmentation performance. AGRNet (Te et al. 2021) introduces an adaptive graph representation approach that learns and reasons over facial components by representing each component as a vertex and relating each vertex, while also incorporating image edges as a prior to refine parsing results. Similarly, EAGRNet (Te et al. 2020) extends this approach by enabling reasoning over non-local regions to capture global dependencies between distinct facial components. Recently, FaRL (Zheng et al. 2022b) explored pre-training on a large image-text face dataset to enhance performance on downstream tasks, demonstrating that their pre-trained weights outperform those based on ImageNet (Deng et al. 2009). DML-CSR (Zheng et al. 2022a) utilizes a multi-task model for face parsing, edge detection, and category edge detection, incorporating a dynamic dual graph convolutional network to address spatial inconsistency and cyclic self-regulation for noisy labels. The recent FP-LIIF (Sarkar et al. 2023) leverages the structural consistency of the human face using a lightweight Local Implicit Function Network with a simple convolutional encoder-pixel decoder architecture, notable for its small parameter size and high FPS, making it ideal for low-compute devices. Despite these advancements, most prior works have focused on specific challenges, such as improving the correlation between facial components, enhancing hair segmentation, handling noisy labels, and optimizing inference speed. However, they often neglect the critical issue of long-tail class performance, leading to sub-optimal results in long-tail classes (see Figure 1).

To overcome this issue, we propose *SegFace*, a systematic approach that enhances the segmentation performance of long-tail classes. These classes are often underrepresented in the dataset, typically including accessories like earring and necklace, while head classes are more frequent and include regions like the face and hair. In a face image, regions like the eyes, mouth, and accessories (long-tail classes) are naturally smaller than the overall face and hair regions (head classes). Using only the final single-scale feature of a model for face segmentation can lead to a loss of detail, as facial features appear at different scales. Our approach leverages a Swin Transformer backbone to extract features at multiple scales, helping to mitigate the scale discrepancy between different face regions. Multi-scale feature extraction effectively captures both fine details and larger structures, aiding the model in capturing the global context of the face. We fuse the multi-scale features using MLP fusion to obtain the fused features, which are then input to the *SegFace* decoder. The lightweight transformer decoder utilizes learnable class-specific tokens, each associated with a particular class. We employ cross-attention between the fused features and learnable tokens, enabling each token to extract class-specific information from the fused features. This design allows the tokens to focus specifically on their corresponding classes, promoting independent modeling of all classes and mitigating the problem of dominant head classes overshadowing long-tail classes during training.

The key contributions of our work are as follows:

- We introduce a lightweight transformer decoder with

learnable class-specific tokens, that ensures each token is dedicated to a specific class, thereby enabling independent modeling of classes. The design effectively addresses the challenge of poor segmentation performance of long-tail classes, prevalent in existing methods.

- Our multi-scale feature extraction and MLP fusion strategy, combined with a transformer decoder that leverages learnable class-specific tokens, mitigates the dominance of head classes during training and enhances the feature representation of long-tail classes.
- *SegFace* establishes a new state-of-the-art performance on the LaPa dataset (93.03 mean F1 score) and the CelebAMask-HQ dataset (88.96 mean F1 score). Moreover, our model can be adapted for fast inference by simply swapping the backbone with a MobileNetV3 backbone. The mobile version achieves a mean F1 score of 87.91 on the CelebAMask-HQ dataset with 95.96 FPS.

2 Related Work

2.1 Face Parsing

Early face parsing approaches employed techniques such as exemplars (Smith et al. 2013), probabilistic index maps (Scheffler and Odobez 2011), Gabor filters (Hernandez-Matamoros et al. 2015), and low-rank decomposition (Guo and Qi 2015). Since the rise of deep learning, numerous deep convolutional network-based methods have been proposed for face segmentation (Warrell and Prince 2009; Khan, Mauro, and Leonardi 2015; Liang et al. 2015; Lin et al. 2019; Liu et al. 2017). Recently, AGRNet (Te et al. 2021) and EAGRNet (Te et al. 2020) proposed graph representation-based methods that correlate different facial components and utilize edge information for parsing. DML-CSR (Zheng et al. 2022a) explores multi-task learning and introduces a dynamic dual graph convolutional network to address spatial inconsistency and cyclic self-regulation to tackle the presence of noisy labels. Local-based methods, which are most similar to our work, aim to predict each facial part individually by training separate models for different facial regions. (Luo, Wang, and Tang 2012) leverages a hierarchical approach to parse each component separately, while (Zhou, Hu, and Zhang 2015) propose using multiple CNNs that take input at different scales, fusing them through an interlinking layer that efficiently integrates local and contextual information. However, existing local-based approaches fail to benefit from a shared backbone and joint optimization, leading to suboptimal performance. *SegFace* addresses this issue by independently modeling all the classes using learnable class-specific tokens, while still benefiting from multi-scale fused features extracted from a shared backbone.

2.2 Transformers

Transformer-based models such as ViT (Dosovitskiy et al. 2020) and DETR (Carion et al. 2020) have demonstrated their effectiveness in segmentation tasks by leveraging attention mechanisms to capture long-range dependencies and global context within images. Segformer (Xie et al. 2021) and SETR (Zheng et al. 2021) are notable works which have

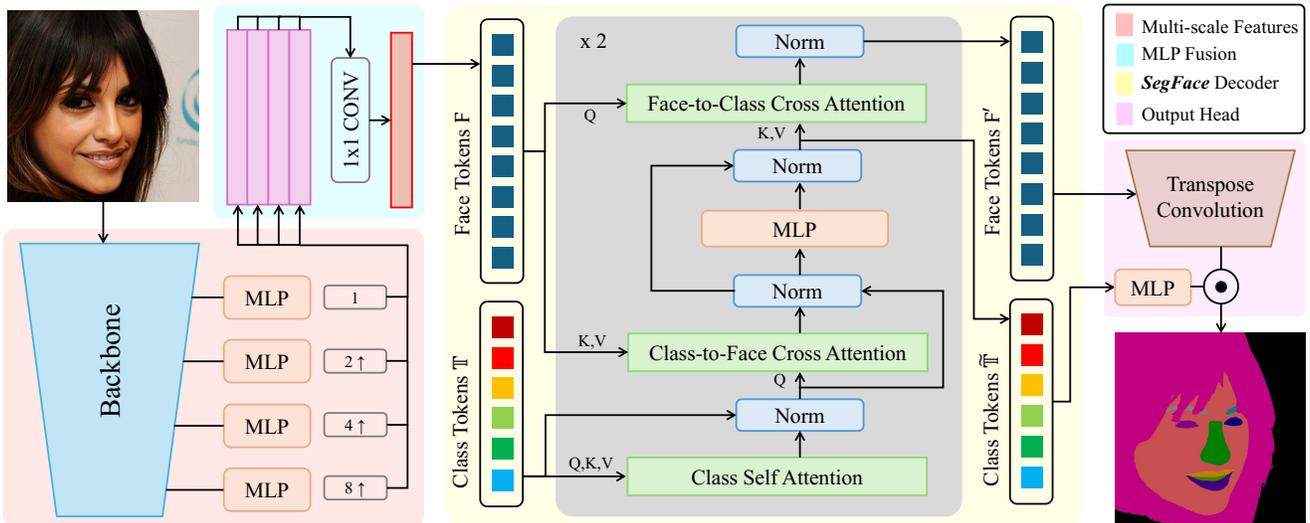


Figure 2: The proposed architecture, *SegFace*, addresses face segmentation by enhancing the performance on long-tail classes through a transformer-based approach. Specifically, multi-scale features are first extracted from an image encoder and then fused using an MLP fusion module to form face tokens. These tokens, along with class-specific tokens, undergo self-attention, face-to-token, and token-to-face cross-attention operations, refining both class and face tokens to enhance class-specific features. Finally, the upscaled face tokens and learned class tokens are combined to produce segmentation maps for each facial region.

shown that transformers can outperform traditional CNNs in general segmentation tasks. However, the application of transformers in face segmentation remains relatively under-explored, despite their potential advantages. Face segmentation presents unique challenges, such as the need for precise boundary detection and sensitivity to subtle variations in facial features, which traditional CNNs have addressed effectively. However, recent transformer-based segmentation networks like Mask2Former (Cheng, Schwing, and Kirillov 2022) and SAM (Kirillov et al. 2023) have shown promising results in capturing both global and fine-grained contexts, leading to more accurate segmentation. These models leverage self-attention and cross-attention mechanisms, which can be viewed as non-local mean operations that compute the weighted average of all inputs. As a result, each class’s inputs are calculated independently and averaged, allowing the model to selectively attend to relevant features without spatial constraints. This leads to a richer, contextualized representation, which can significantly benefit the understanding of long-tail visual relationships.

3 Proposed Work

The human face consists of various regions, including the nose, eyes, mouth, and accessories like earrings and necklaces. In face segmentation, these regions are treated as different classes, which vary in scale and frequency of occurrence. Classes such as hair and nose, naturally appear more often in a face image and are referred to as head classes. In contrast, accessories, which may not be present in every face image, are called long-tail classes and are underrepresented in face segmentation datasets. We calculate the frequency of each class in the dataset and determine the probability of a class occurring in a face image of the CelebAMask-HQ

dataset. As shown in Figure 1, the probability of a head class being present in an image is approximately 1.0, compared to 0.26 and 0.05 for long-tail classes. Upon analyzing current face segmentation methods, we observe that they often perform poorly on long-tail classes. Our goal is to enhance the segmentation performance of long-tail classes, thereby boosting overall face segmentation performance.

Given a batch of face images $I \in \mathbb{R}^{B \times H \times W \times 3}$, consisting of N classes, where B is the batch size, while H and W denote the height and width of the image, respectively. *SegFace* extracts multi-scale features $\mathbb{G} = \{G_i | 1 \leq i \leq 4\}$ from the intermediate layers of the image encoder E_θ . These features are then fused using a MLP fusion module f_ϕ to obtain the face tokens F . The face tokens, along with their corresponding positional encodings, and the learnable class-specific tokens $\mathbb{T} = \{T_i | 1 \leq i \leq N\}$, are processed by the light-weight *SegFace* decoder g_ψ through self-attention and cross-attention operations, resulting in the learned class tokens $\tilde{\mathbb{T}}$ and updated face tokens F' . The updated face tokens are then upscaled using an upscaling module h_α and multiplied element-wise with the learned class tokens $\tilde{\mathbb{T}}$ after the tokens has been passed through an MLP to obtain the final segmentation map $\mathbb{S} = \{S_i | 1 \leq i \leq N\}$ where $S_i \in \mathbb{R}^{B \times 1 \times H \times W}$, represents the segmentation map for each class. The complete process is as follows:

$$\tilde{\mathbb{T}}, F' = g_\psi(F, \mathbb{T}), \quad \text{where } F = f_\phi(E_\theta(I))$$

$$S_i = h_\alpha(F') \odot \text{MLP}(\tilde{\mathbb{T}}_i)$$

Here, S_i is the output segmentation map for the i -th class. We utilize these segmentation maps to calculate the loss. We use cross entropy loss along with dice loss to train the complete pipeline which is illustrated in Figure 2. The final loss function can be given as: $L = \lambda_1 L_{\text{dice}} + \lambda_2 L_{\text{CE}}$.

3.1 Multi-scale Feature Extraction

We perform multi-scale feature extraction to address the problem of scale discrepancy between different face regions. This approach effectively captures both fine details and larger structures, helping to obtain a comprehensive global context of the face and better handle the varying sizes and shapes of facial components. The multi-scale features are extracted from the image encoder E_θ . Let the batch of input images be $I \in \mathbb{R}^{B \times H \times W \times 3}$, where B is the batch size, and H and W are the height and width of the image. The encoder extracts features from multiple layers:

$$\mathbb{G} = \{G_i \mid 1 \leq i \leq 4\}, \quad G_i \in \mathbb{R}^{B \times C_i \times H_i \times W_i}$$

Here, G_i represents the feature map extracted from the i -th layer of the encoder, C_i is the number of channels in the i -th feature map, and H_i and W_i denote the height and width of the i -th feature map, respectively. The hierarchical features extracted from the encoder help capture coarse to fine-grained representations, making them suitable for segmenting smaller classes, which are often long-tail classes.

3.2 MLP Fusion

We perform multi-scale feature aggregation using an MLP fusion module f_θ to obtain the face tokens that will be passed to the *SegFace* decoder. In this module, the multi-scale features $\mathbb{G} = \{G_i \mid 1 \leq i \leq 4\}$ are processed by separate MLPs, each corresponding to a different scale, to make the channel dimension consistent for fusion. Each MLP transforms its corresponding G_i into a feature map G'_i with a uniform number of channels C' , as follows: $G'_i = \text{MLP}_i(G_i)$, where $G'_i \in \mathbb{R}^{B \times C' \times H_i \times W_i}$. The resulting feature maps G'_i are then upsampled to match the spatial resolution of the first feature map G'_1 using bilinear interpolation, represented as $G''_i = \text{Interp}(G'_i)$, where $G''_i \in \mathbb{R}^{B \times C' \times H_1 \times W_1}$, $\forall i \in \{1, 2, 3, 4\}$. These upsampled multi-scale features G''_i are concatenated along the channel dimension to form a unified feature map. Finally, this concatenated feature map is passed through a single convolutional layer, to reduce the channel dimensionality back to C' .

$$F_{\text{concat}} = \text{Concat}(G''_1, G''_2, G''_3, G''_4) \in \mathbb{R}^{B \times (4 \times C') \times H_1 \times W_1}$$

$$F = \text{Conv1x1}(F_{\text{concat}}) \in \mathbb{R}^{B \times C' \times H_1 \times W_1}$$

This fused feature map F represents the final multi-scale face tokens which is given as input to the *SegFace* decoder.

3.3 SegFace Decoder

The *SegFace* decoder is designed to model each class independently while enabling interactions between them, using learnable class-specific tokens. Let $\mathbb{T} = T_i \in \mathbb{R}^{1 \times D} \mid 1 \leq i \leq N$ represent these tokens, where N is the number of classes, and D is the embedding dimension (here, $D = 256$). These tokens are appended with positional encodings and correspond to various facial components, such as the background, face, eyes, nose, and other features. The decoder comprises of three main components: 1) Class-token Self-Attention, 2) Class-token to Face-token Cross-Attention, and 3) Face-token to

Class-token Cross-Attention. Through self-attention and cross-attention operations within the transformer decoder, the tokens are guided to focus on class-specific features and facilitate interaction among different facial regions.

Class-token Self-Attention: This component facilitates interaction between different regions of the face by allowing each class token, T_i , to attend to all other class tokens. For each class token T_i , the operation is defined as:

$$T'_i = \text{SelfAttention}(Q = T_i, K = \mathbb{T}, V = \mathbb{T}),$$

where SelfAttention denotes the multi-head self-attention operation, and Q , K , and V represent the queries, keys, and values, respectively. Each class token corresponds to a specific class, and the SelfAttention operation enables the model to learn the correlations between the structure and position of different facial regions.

Class-token to Face-token Cross-Attention: In this component, each class token T'_i attends to the fused face token F , facilitating the extraction of class-specific information and enabling independent modeling of the classes. The updated class token \tilde{T}_i is computed as follows:

$$\tilde{T}_i = \text{CrossAttention}(Q = T'_i, K = F, V = F),$$

where CrossAttention denotes the cross-attention operation. This mechanism ensures that long-tail classes are not overshadowed during training, as each class is associated with a token that extracts relevant features specifically for segmenting that long-tail class.

Face-token to Class-token Cross-Attention: In this component, the fused face tokens attend back to the learned class tokens, refining the face representation with class-specific information. The refined face token F' is computed as follows:

$$F' = \text{CrossAttention}(Q = F, K = \tilde{\mathbb{T}}, V = \tilde{\mathbb{T}})$$

This component guides the feature extraction and fusion modules by aligning their training to ensure that the extracted features are enriched with class-specific information.

3.4 Output Head

The output head's role is to generate the final segmentation maps from the learned class-specific tokens and the updated face tokens. The face tokens F' are upsampled using a small network h_α , which comprises transpose convolution operations. The upscaling increases the resolution of the face tokens to match the original image size. Formally, this can be defined as $U = h_\alpha(F')$, where $U \in \mathbb{R}^{B \times C' \times H \times W}$ is the upsampled face token embedding, and C' is the reduced embedding dimension after upscaling. Finally, the learned class-specific tokens $\tilde{\mathbb{T}} = \{\tilde{T}_i \mid 1 \leq i \leq N\}$ are passed through an MLP and then multiplied element-wise with the upsampled face tokens to produce the final segmentation maps:

$$S_i = U \odot \text{MLP}(\tilde{T}_i),$$

where \odot denotes element-wise multiplication, and $S_i \in \mathbb{R}^{B \times 1 \times H \times W}$ represents the segmentation map for the i -th class. The final output is a set of segmentation maps $\mathbb{S} = \{S_i \mid 1 \leq i \leq N\}$ for all classes, where each S_i corresponds to a specific face component, effectively segmenting the input face image into its respective regions.

Method	Venue	Resolution	Skin	Hair	Nose	L-Eye	R-Eye	L-Brow	R-Brow	L-Lip	I-Mouth	U-Lip	Mean F1 \uparrow	Mean IoU \uparrow
Wei et al.	TIP 19	512	96.1	95.1	96.1	88.9	87.5	86.0	87.8	83.8	89.2	83.1	89.36	-
BASS	AAAI 20	473	97.2	96.3	95.5	88.1	88.0	87.7	87.6	85.7	87.6	84.4	89.81	-
EAGRNet	ECCV 20	473	97.3	96.2	97.1	89.5	90.0	86.5	87.0	89.0	90.0	88.1	91.07	-
AGRNet	TIP 21	473	97.7	96.5	97.3	91.6	91.1	89.9	90.0	90.1	90.7	88.5	92.34	-
FaRL _{scratch}	CVPR 22	512	97.2	93.1	97.3	91.6	91.5	90.1	89.7	89.1	89.4	87.2	91.62	-
DML-CSR	CVPR 22	473	97.6	96.4	97.3	91.8	91.5	90.4	90.4	89.9	90.5	88.0	92.38	87.13
FP-LIIF	CVPR 23	512	97.5	95.9	97.2	92.0	92.2	90.9	90.6	89.5	90.3	87.7	92.38	-
<i>SegFace</i>	AAAI 25	224	97.5	95.4	97.3	91.9	92.1	90.9	90.8	89.9	90.8	88.3	92.50	87.26
<i>SegFace</i>	AAAI 25	256	97.5	95.7	97.3	92.2	92.2	91.0	90.8	90.0	91.0	88.4	92.61	87.45
<i>SegFace</i>	AAAI 25	448	97.7	96.2	97.5	92.6	92.7	91.6	91.4	90.5	91.4	88.8	93.03	88.13
<i>SegFace</i>	AAAI 25	512	97.7	96.3	97.5	92.6	92.7	91.6	91.4	90.5	91.2	88.7	93.03	88.14

(a) LaPa Dataset

Method	Venue	Resolution	Face	Nose	E-Glasses	L-Eye	R-Eye	L-Brow	R-Brow	L-Ear	R-Ear	Mean F1 \uparrow	Mean IoU \uparrow
			I-Mouth	U-Lip	L-Lip	Hair	Hat	Earring	Necklace	Neck	Cloth		
Wei et al.	TIP 19	512	96.4	91.9	89.5	87.1	85.0	80.8	82.5	84.1	83.3	82.06	-
			90.6	87.9	91.0	91.1	83.9	65.4	17.8	88.1	80.6		
EAGRNet	ECCV 20	473	96.2	94.0	92.3	88.6	89.0	85.7	85.2	88.0	85.7	84.89	-
			95.0	88.9	91.2	94.9	82.7	68.3	27.6	89.4	85.3		
AGRNet	TIP 21	473	96.5	93.9	91.8	88.7	89.1	85.5	85.6	88.1	88.7	85.12	-
			92.0	89.1	91.1	87.6	87.2	69.6	32.8	89.9	84.9		
FaRL _{scratch}	CVPR 22	512	96.2	93.8	92.3	89.0	89.0	85.3	85.4	86.9	87.3	84.77	-
			91.7	88.1	90.0	94.9	82.7	63.1	33.5	90.8	85.9		
DML-CSR	CVPR 22	473	95.7	93.9	92.6	89.4	89.6	85.5	85.7	88.3	88.2	86.07	77.81
			91.8	89.1	91.0	94.5	88.5	69.6	40.6	89.6	85.7		
FP-LIIF	CVPR 23	512	96.6	94.0	92.5	90.0	90.1	85.6	85.4	86.8	86.7	86.14	-
			92.7	89.4	91.3	95.2	86.7	67.2	42.2	91.4	86.8		
<i>SegFace</i>	AAAI 25	224	96.4	93.8	94.0	90.1	90.2	86.0	86.0	88.2	87.5	87.47	79.65
			92.2	89.4	90.7	95.7	89.6	71.1	52.6	91.5	89.5		
<i>SegFace</i>	AAAI 25	256	96.5	93.9	94.3	90.2	90.5	86.3	86.4	88.5	88.0	87.66	79.91
			92.4	89.6	90.9	95.8	89.7	72.0	52.8	91.5	88.7		
<i>SegFace</i>	AAAI 25	448	96.6	94.1	95.0	90.8	90.9	87.0	86.9	89.2	88.6	88.77	81.30
			92.9	90.0	91.3	96.0	89.9	74.5	62.0	92.0	90.0		
<i>SegFace</i>	AAAI 25	512	96.7	94.2	95.4	90.9	91.1	87.2	87.1	89.3	88.9	88.96	81.55
			93.1	90.3	91.6	96.0	89.3	75.1	63.4	92.1	89.8		

(b) CelebAMask-HQ dataset

Table 1: Quantitative results on (a) LaPa dataset and (b) CelebAMask-HQ dataset

4 Experiments

4.1 Datasets

We conduct our experiments on three standard face segmentation datasets: LaPa (Liu et al. 2020), CelebAMask-HQ (Lee et al. 2020b), and Helen (Le et al. 2012). The LaPa dataset contains a total of 22,168 images, with 18,176 used for training, 2,000 for validation, and 2,000 for testing. This dataset is annotated for 11 classes, including skin, hair, nose, left eye, right eye, left brow, right brow, upper lip, and lower lip. The CelebAMask-HQ dataset comprises 30,000 face images, split into 24,183 for training, 2,993 for validation, and 2,824 for testing. It features 19 semantic classes, including accessories such as earring, necklace, eyeglass, and hat, which are considered long-tail classes due to their infrequent occurrence in the dataset. The other classes are the same as those in the LaPa dataset, with the addition of left/right ear, cloth and neck. The Helen dataset, being the smallest, consists of 2,000 training samples, 230 validation samples, and 100 test samples, annotated for 11 classes.

4.2 Implementation Details

We trained *SegFace* in various configurations by changing the backbones (Swin, Swin V2, ResNet101, MobileNetV3, EfficientNet) and input resolutions (64, 96, 128, 192, 224, 256, 448, 512). The models were optimized for 300 epochs using the AdamW optimizer, with an initial learning rate of $1e^{-4}$ and a weight decay of $1e^{-5}$. We employed a step LR scheduler with a gamma value of 0.1, which reduces the learning rate by a factor of 0.1 at epochs 80 and 200. A batch size of 32 was used for training on the LaPa and CelebAMask-HQ datasets, and 16 for the Helen dataset. We did not perform any augmentations on the CelebAMask-HQ and Helen datasets. For the LaPa dataset, we applied random rotation $[-30^\circ, 30^\circ]$, random scaling $[0.5, 3]$, and random translation $[-20\text{px}, 20\text{px}]$, along with RoI tanh warping (Lin et al. 2019) to ensure that the network focused on the face region. The λ_1 and λ_2 values were set at 0.5 for dice loss and cross entropy loss, respectively. Our method was evaluated against other baselines using class-wise F1 score, mean F1 score, and mean IoU, with the background class excluded in all metrics. All code was implemented in PyTorch, and the

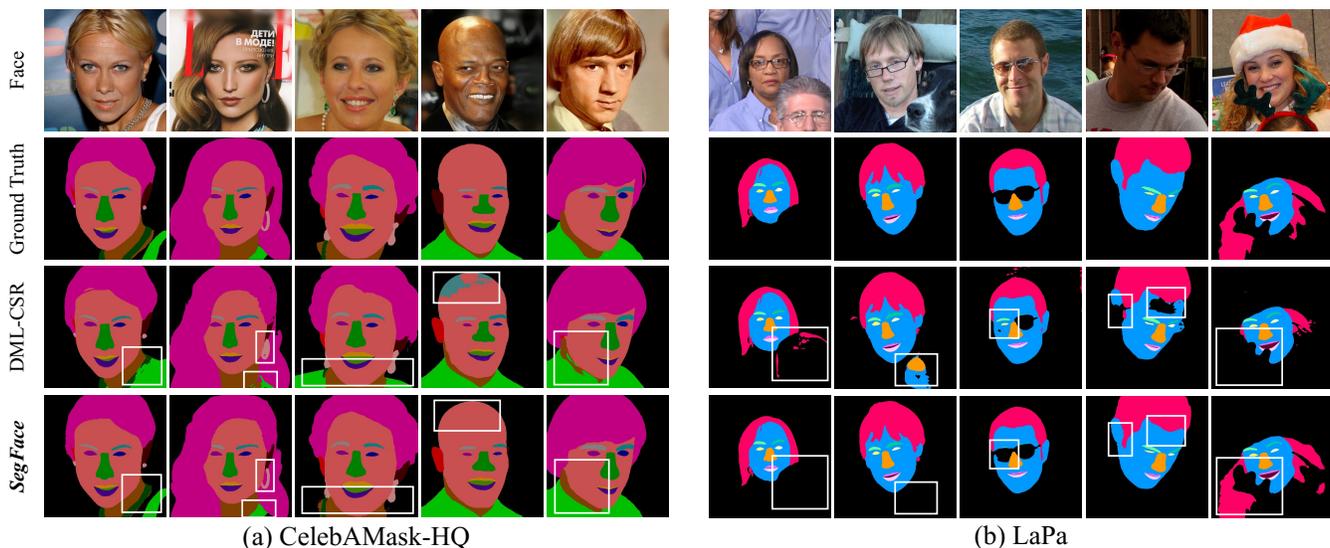


Figure 3: The qualitative comparison highlights the superior performance of our method, *SegFace*, compared to DML-CSR. In (a), *SegFace* effectively segments both long-tail classes like earrings and necklaces as well as head classes such as hair and neck. In (b), it also excels in challenging scenarios involving multiple faces, human-resembling features, poor lighting, and occlusion, where DML-CSR struggles.

models were trained on eight A6000 GPUs, each equipped with 48 GB of memory.

5 Results and Analysis

In this section, we detail the quantitative and qualitative results of *SegFace* and demonstrate its superiority in handling the segmentation of long-tail classes. Further, we analyze the benefits of the proposed method.

Quantitative Results: The class-wise F1-score, mean F1-score, and mean IoU on the LaPa and CelebAMask-HQ datasets are shown in Table 1(a) and Table 1(b), respectively. We observe that *SegFace* outperforms other existing methods, achieving a mean F1-score of 93.03 and a mean IoU of 88.14 on the LaPa dataset. We see improvements in majority of the classes, with the largest gains in the lower-lip, inner-mouth, and upper-lip classes, with increments of 0.6, 0.7, and 0.7, respectively. The performance improvement in these classes validates our claim that multi-scale feature extraction and fusion help mitigate the scale-discrepancy problem between different facial regions, thereby boosting overall segmentation performance. *SegFace* also significantly outperforms other baselines on the CelebAMask-HQ dataset, achieving a mean F1-score of 88.96 (+2.89) and a mean IoU of 81.55 (+3.74). Specifically, we observe significant improvements in the long-tail classes such as eyeglasses, earrings, and necklaces, with increments of 2.8, 5.5, and 22.8, respectively. In addition to these improvements in long-tail classes, *SegFace* also shows enhanced performance across other classes in the CelebAMask-HQ dataset, outperforming other methods when comparing the class-wise F1 score. This significant performance improvement can be attributed to the transformer decoder with learnable class-specific tokens. It associates each class with a specific token

and prevents the dominance of head classes during training, ensuring effective feature representation for the long-tail classes. Additionally, the cross-attention between fused features and tokens helps the tokens extract class-specific information and enables independent modeling of classes.

Qualitative Results: We illustrate the qualitative comparison of our proposed method against other baselines in Figure 3. From Figure 3(a) [columns 1,2,3], we validate that *SegFace* is capable of segmenting long-tail classes such as earring and necklace much better compared to the existing state-of-the-art method, DML-CSR. This demonstrates the effectiveness of the proposed transformer decoder with learnable task-specific queries. It enables independent modeling of all classes by associating each token with a particular class. In this design, the token can focus specifically on that class and learn to leverage the fused features for segmentation. Furthermore, from Figure 3(a) [columns 4,5], we observe that the proposed method also performs better on head classes such as hair and neck. The results on the LaPa dataset, as shown in Figure 3(b) [columns 1, 2], indicate that DML-CSR struggles with face segmentation in the presence of multiple faces or human-resembling features in the vicinity. We mitigate this issue by incorporating RoI Tanh warping (Lin et al. 2019) to ensure that the model focuses on the face region while performing segmentation. From Figure 3(b) [columns 3,4], we can see that DML-CSR performs poorly in challenging lighting conditions and in Figure 3(b) [column 5], it struggles with occlusion. *SegFace* outperforms DML-CSR and is able to accurately segment facial regions even in these complex scenarios.

Analysis: We make the following claims: “The transformer decoder with learnable class-specific queries enables independent modeling of classes” and “In our proposed ap-

proach, each token is associated with one class, allowing it to focus specifically on that particular class.” To validate these claims, we analyze what each token is learning. We visualize the segmentation outputs of some tokens such as upper-lip, nose, left-brow and right-eye in Figure 4(a). We observe that each token effectively learns the class it has been associated with, demonstrating independent modeling of classes. The learnable tokens leverage the shared fused features via cross-attention to learn the class-specific information. Furthermore, we manually analyzed the segmentation outputs and compared them with the ground truth. We found that the proposed approach provides accurate segmentation output even in the presence of samples with noisy ground truths, showcasing its robustness. The noisy ground truths and our predictions for the same are illustrated in Figure 4(b).



(a)



(b)

Figure 4: (a) Class-specific tokens segment their corresponding classes, showcasing the independent modeling of each class. (b) Comparison of noisy ground truth with prediction from *SegFace*

6 Ablation Studies

We conduct an ablation analysis to study different components in our proposed approach and provide helpful insights.

Varying the backbone of *SegFace*: We trained *SegFace* with various backbones to demonstrate the strength of the proposed lightweight transformer decoder with learnable task-specific tokens. As shown in Table 2(a), we conducted experiments using backbones with parameter sizes ranging from 7M to 91M and observed that the segmentation per-

Backbone	Mean F1	Mean IoU	FPS	Params
ResNet 100	87.50	79.65	67.81	47.254
EfficientNet	88.94	81.49	40.14	57.035
MobileNet V3	87.91	79.98	95.96	7.034
Swin V2	88.73	81.30	34.13	91.168
Swin (w/o fusion)	87.83	80.07	40.00	90.513
Swin	88.96	81.55	38.95	91.006

(a) *SegFace* performance with different backbones

Res	64	96	128	192	224	256	448	512
FPS	54.56	54.11	45.77	47.39	47.72	42.78	44.53	38.95
mF1	80.92	83.75	85.62	87.11	87.47	87.66	88.77	88.96
mIoU	71.72	75.20	77.24	79.18	79.65	79.91	81.30	81.55

(b) *SegFace* performance for varying image resolutions

Table 2: Ablation study for different backbones and varying image resolution.

formance remained consistent with minimal variation. This consistency indicates that the transformer decoder is responsible for majority of the heavy lifting, making it the core component of our proposed approach. Furthermore, we want to emphasize that the proposed method can be adapted for low-compute edge devices by simply swapping the backbone to MobileNetV3 (Howard et al. 2019). The mobile version achieves 95.96 FPS with a mean F1 score of 87.91 (+1.77) on the CelebAMask-HQ dataset, surpassing the current state-of-the-art.

***SegFace* w/o multi-scale feature extraction:** We trained *SegFace* using the single-scale final feature obtained from the backbone without any feature fusion, as shown in Table 2(a) [Row 5]. As expected, we observed a drop in performance when the model was trained without multi-scale feature extraction. This showcases the importance of multi-scale feature extraction and feature fusion in effectively handling different face regions that appear at varying scales.

Performance at different input resolutions: We analyzed the performance and FPS of *SegFace* at different input resolutions to showcase the trade-off between FPS and performance, which can be valuable for applications requiring lower memory usage and inference costs. Notably, *SegFace*, even when trained at a low-resolution of 192×192 , outperforms the best version of current state-of-the-art DML-CSR, which is trained at 512×512 resolution.

7 Conclusion

In this work, we present *SegFace*, a systematic approach that leverages a lightweight transformer decoder with learnable task-specific tokens to address the challenge of poor segmentation performance on long-tail classes. We also incorporate multi-scale feature extraction and MLP fusion in our pipeline to resolve the scale discrepancy problem between different face regions. Through extensive experiments, we validate the effectiveness of our approach and provide insightful comments to highlight its superiority. The results demonstrate that we significantly outperform other methods, achieving state-of-the-art segmentation performance on the LaPa and CelebAMask-HQ datasets.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, 213–229. Springer.
- Cheng, B.; Schwing, A.; and Kirillov, A. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. *arXiv preprint arXiv:2208.02717*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.
- Guo, R.; and Qi, H. 2015. Facial feature parsing and landmark detection via low-rank matrix decomposition. In *2015 IEEE International Conference on Image Processing (ICIP)*, 3773–3777. IEEE.
- Guo, T.; Kim, Y.; Zhang, H.; Qian, D.; Yoo, B.; Xu, J.; Zou, D.; Han, J.-J.; and Choi, C. 2018. Residual encoder decoder network and adaptive prior for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32-1.
- Hernandez-Matamoros, A.; Bonarini, A.; Escamilla-Hernandez, E.; Nakano-Miyatake, M.; and Perez-Meana, H. 2015. A facial expression recognition with automatic segmentation of face regions. In *Intelligent Software Methodologies, Tools and Techniques: 14th International Conference, SoMet 2015, Naples, Italy, September 15-17, 2015. Proceedings 14*, 529–540. Springer.
- Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; Le, Q. V.; and Adam, H. 2019. Searching for MobileNetV3. *CoRR*, abs/1905.02244.
- Khan, K.; Mauro, M.; and Leonardi, R. 2015. Multi-class semantic segmentation of faces. In *2015 IEEE International Conference on Image Processing (ICIP)*, 827–831. IEEE.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, P.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643*.
- Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; and Huang, T. S. 2012. Interactive facial feature localization. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, 679–692. Springer.
- Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020a. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5549–5558.
- Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020b. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Li, Y.; Liu, S.; Yang, J.; and Yang, M.-H. 2017. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3911–3919.
- Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; and Yan, S. 2015. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, 37(12): 2402–2414.
- Lin, J.; Yang, H.; Chen, D.; Zeng, M.; Wen, F.; and Yuan, L. 2019. Face parsing with roi tanh-warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5654–5663.
- Lin, Y.; Shen, J.; Wang, Y.; and Pantic, M. 2021. Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, 112: 104190.
- Liu, S.; Shi, J.; Liang, J.; and Yang, M.-H. 2017. Face parsing via recurrent propagation. *arXiv preprint arXiv:1708.01936*.
- Liu, Y.; Shi, H.; Shen, H.; Si, Y.; Wang, X.; and Mei, T. 2020. A new dataset and boundary-attention semantic segmentation for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34-07, 11637–11644.
- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.
- Luo, P.; Wang, X.; and Tang, X. 2012. Hierarchical face parsing via deep learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2480–2487. IEEE.
- Sarkar, M.; Nikitha, S.; Hemani, M.; Jain, R.; and Krishnamurthy, B. 2023. Parameter Efficient Local Implicit Image Function Network for Face Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20970–20980.
- Scheffler, C.; and Odobez, J.-M. 2011. Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps. In *Proceedings of the British Machine Vision Conference*, 53–1.
- Smith, B. M.; Zhang, L.; Brandt, J.; Lin, Z.; and Yang, J. 2013. Exemplar-based face parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3484–3491.

- Te, G.; Hu, W.; Liu, Y.; Shi, H.; and Mei, T. 2021. Agr-net: Adaptive graph representation learning and reasoning for face parsing. *IEEE Transactions on Image Processing*, 30: 8236–8250.
- Te, G.; Liu, Y.; Hu, W.; Shi, H.; and Mei, T. 2020. Edge-aware graph representation learning and reasoning for face parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 258–274. Springer.
- Wan, Z.; Chen, H.; An, J.; Jiang, W.; Yao, C.; and Luo, J. 2022. Facial attribute transformers for precise and robust makeup transfer. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1717–1726.
- Warrell, J.; and Prince, S. J. 2009. Labelfaces: Parsing facial features by multiclass labeling with an epitome prior. In *2009 16th IEEE international conference on image processing (ICIP)*, 2481–2484. IEEE.
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv preprint arXiv:2105.15203*.
- Xu, C.; Zhang, J.; Hua, M.; He, Q.; Yi, Z.; and Liu, Y. 2022. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7632–7641.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.
- Zheng, Q.; Deng, J.; Zhu, Z.; Li, Y.; and Zafeiriou, S. 2022a. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4156–4165.
- Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; and Torr, P. H. 2021. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6881–6890.
- Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022b. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18697–18709.
- Zhou, Y.; Hu, X.; and Zhang, B. 2015. Interlinked convolutional neural networks for face parsing. In *Advances in Neural Networks–ISNN 2015: 12th International Symposium on Neural Networks, ISNN 2015, Jeju, South Korea, October 15–18, 2015, Proceedings 12*, 222–231. Springer.