Parameter Efficient Local Implicit Image Function Network for Face Segmentation

Mausoom Sarkar
MDSR Lab, Adobe

Nikitha SR*
IIT Madras

Mayur Hemani
MDSR Lab, Adobe

Rishabh Jain
MDSR Lab, Adobe

Balaji Krishnamurthy
MDSR Lab, Adobe

Abstract

Face parsing is defined as the per-pixel labeling of images containing human faces. The labels are defined to identify key facial regions like eyes, lips, nose, hair, etc. In this work, we make use of the structural consistency of the human face to propose a lightweight face-parsing method using a Local Implicit Function network, FP-LIIF. We propose a simple architecture having a convolutional encoder and a pixel MLP decoder that uses $1/26^{th}$ number of parameters compared to the state-of-the-art models and yet matches or outperforms state-of-the-art models on multiple datasets, like CelebAMask-HQ and LaPa. We do not use any pretraining, and compared to other works, our network can also generate segmentation at different resolutions without any changes in the input resolution. This work enables the use of facial segmentation on low-compute or low-bandwidth devices because of its higher FPS and smaller model size.

1. Introduction

Face parsing is the task of assigning pixel-wise labels to a face image to distinguish various parts of a face, like eyes, nose, lips, ears, etc. This segregation of a face image enables many use cases, such as face image editing [20, 46, 57], face e-beautification [37], face swapping [16, 35, 36], face completion [23].

Since the advent of semantic segmentation through the use of deep convolutional networks [31], a multitude of research has investigated face parsing as a segmentation problem through the use of fully convolutional networks [13, 14, 25, 26, 28, 29]. In order to achieve better results, some methods [14, 28] make use of conditional random fields (CRFs), in addition to CNNs. Other methods [25, 27],
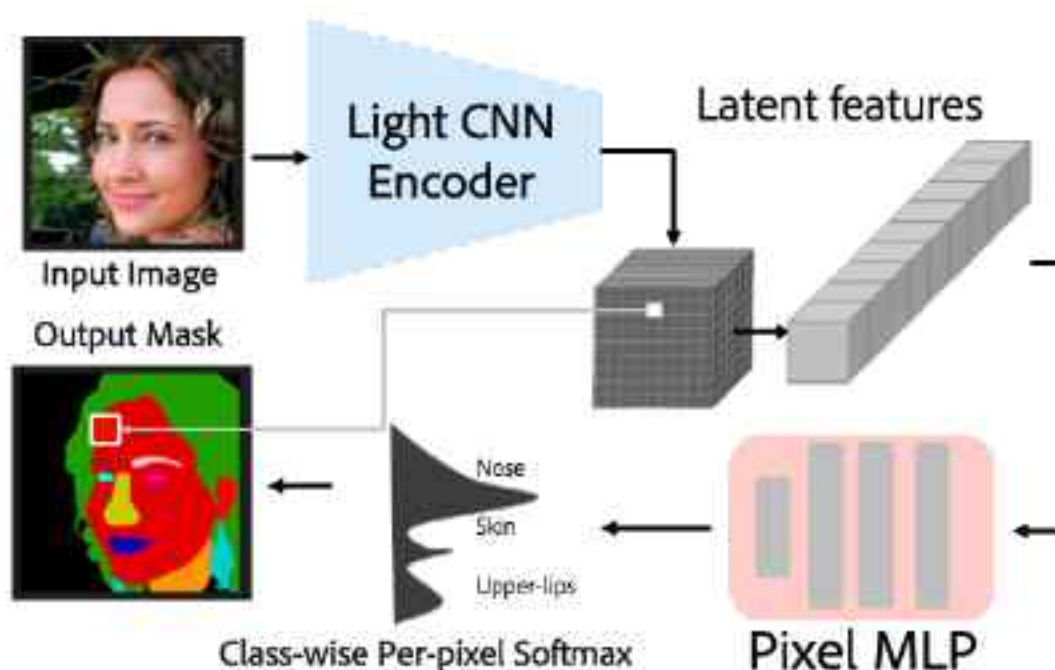

Figure 1. The simple architecture of Local Implicit Image representation base FP-LIIF: A light convolutional encoder of modified resblocks followed by a pixel only MLP decoder

focus on a two-step approach that predicts bounding boxes of facial regions (nose, eyes, hair. etc.) followed by segmentation within the extracted regions. Later works like AGRNET [48] and EAGR [49] claim that earlier approaches do not model the relationship between facial components and that a graph-based system can model these statistics, leading to more accurate segmentation.

In more recent research, works such as FaRL [59] investigate pretraining on a human face captioning dataset. They pre-train a Vision Transformer (ViT) [8] and finetune on face parsing datasets and show improvement in comparison to pre-training with classification based pre-training like ImageNet [44], etc., or no pre-training at all. The current state-of-the-art model, DML_CSR [58], tackles the face parsing task using multiple concurrent strategies including multi-task learning, graph convolutional network (GCN), and cyclic learning. The Multi-task approach handles edge discovery in addition to face segmentation. The proposed

---

*Work done during internship at Adobe

GCN is used to provide global context instead of an average pooling layer. Additionally, cyclic learning is carried out to arrive at an ensemble model and subsequently perform self-distillation using the ensemble model in order to learn in the presence of noisy labels.

In this work, we perform face segmentation by taking advantage of the consistency seen in human facial structures. We take our inspiration from various face modeling works [1, 12, 61] that can reconstruct a 3D model of a face from 2D face images. These works show it is possible to create a low-dimensional parametric model of the human face in 3D. This led us to conclude that 2D modeling of the human face should also be possible with low dimension parametric model. Recent approaches, like NeRF [34] and Siren [47] demonstrated that it is possible to reconstruct complex 3D and 2D scenes with implicit neural representation. Many other works [2, 11, 43, 56] demonstrate that implicit neural representation can also model faces both in 3D and 2D. However, to map 3D and 2D coordinates to the RGB space, the Nerf [34] and Siren [47] variants of the models require training a separate network for every scene. This is different from our needs, one of which is that we must map an RGB image into label space and require a single network for the whole domain. That brings us to another method known as LIIF [3], which is an acronym for a Local Implicit Image Function and is used to perform image super-resolution. They learn an approximation of a continuous function that can take in any RGB image with low resolution and output RGB values at the sub-pixel level. This allows them to produce an enlarged version of the input image. Thus, given the current success of learning implicit representations and the fact that the human face could be modeled using a low-dimension parametric model, we came to the conclusion that a low parameter count LIIF-inspired model should learn a mapping from a face image to its label space or segmentation domain. In order to test this hypothesis, we modify a low-parameter version of EDSR [24] encoder such that it can preserve details during encoding. We also modify the MLP decoder to reduce the computing cost of our decoder. Finally, we generate a probability distribution in the label space instead of RGB values. We use the traditional cross-entropy-based losses without any complicated training mechanisms or loss adaptations. An overview of the architecture is depicted in Figure 1, and more details are in Section 3. Even with a parameter count that is $1/26^{th}$ compared to DML_CSR [58], our model attains state-of-the-art F1 and IoU results for CelebAMask-HQ [21] and LaPa [29] datasets. Some visualizations of our outputs are shared in Figure 3 and Figure 4.

To summarise, our key contributions are as follows:

- We propose an implicit representation-based simple and lightweight neural architecture for human face semantic segmentation.

- We establish new state-of-the-art mean F1 and mean IoU scores on CelebAMask-HQ [21] and LaPa [29].
- Our proposed model has a parameter count of $1/26^{th}$ or lesser compared to the previous state-of-the-art model. Our model's SOTA configuration achieves an FPS of 110 compared to DML_CSR's FPS of 76.

2. Related Work

2.1. Face Parsing

Since face parsing intrinsically involves capturing the parametric relationship between the facial regions, the existing methods in face parsing aim at modeling the spatial dependencies existing in the pixels of the image. Multiple deep learning-based models with multi-objective frameworks have been proposed to handle spatial or inter-part correlations and boundary inconsistencies and capture the image's global context. Liu et al. [29] proposed a two-head CNN that uses an encoder-decoder framework with spatial pyramid pooling to capture global context. The other head uses the shared features from the encoder to predict a binary map of confidence that a given pixel is a boundary which is later combined with the features from the first head to perform face parsing. EAGRnet [49] uses graph convolution layers to encode spatial correlations between face regions into the vertices of a graph. Zheng et al. [58] combine these approaches to build DML-CSR, a dual graph convolution network that combines graph representations obtained from shallow and deeper encoder layers to capture global context. Additionally, they employ multi-task learning by adding edge and boundary detection tasks and weighted loss functions to handle these tasks. They also use an ensemble-based distillation training methodology claiming that it helps in learning in the presence of noisy labels. They achieve state-of-the-art performance on multiple face-parsing datasets. Recently a transformer-based approach has also achieved state-of-the-art performance but with the help of training with additional data. FaRL [59] starts by pre-training a model with a face-image captioning dataset to learn an encoding for face images and their corresponding captions. Their image encoder, a ViT [8], and a transformer-based text encoder from CLIP [42] learn a common feature encoding using contrastive learning. They then use the pre-trained image encoder and finetune on various face-related task datasets to report state-of-the-art numbers. We compare our performance numbers with a non-pre-trained version of FaRL [59] because we wanted to test our model on only the task-related dataset, and using additional image-caption data was out-of-the scope of this work.

While these approaches handle the image as a whole to predict a single mask segmenting all the components
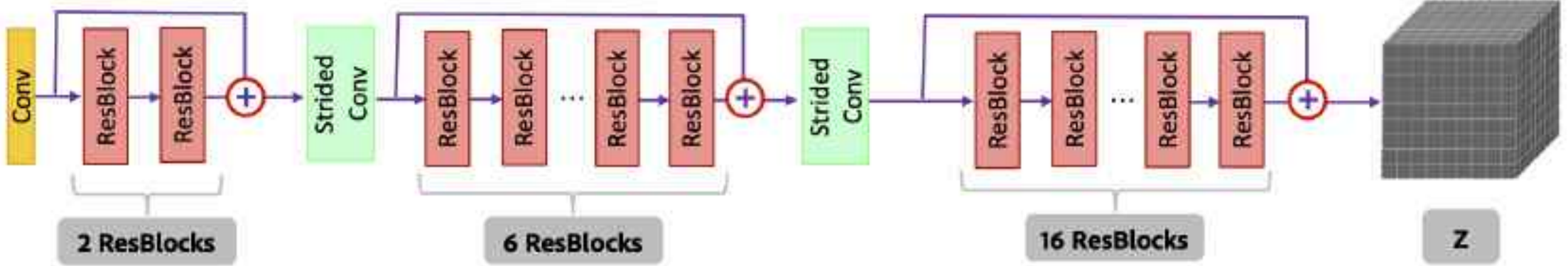
Figure 2. Encoder Architecture: It has three res-block groups. The first two (2,6) res-block groups, followed by a strided convolution per group, are mainly used to reduce the spatial dimensions of the activation maps. The final group of res-blocks creates the grid of features vectors Z. Notice each res-block group has a group-level residual connection.
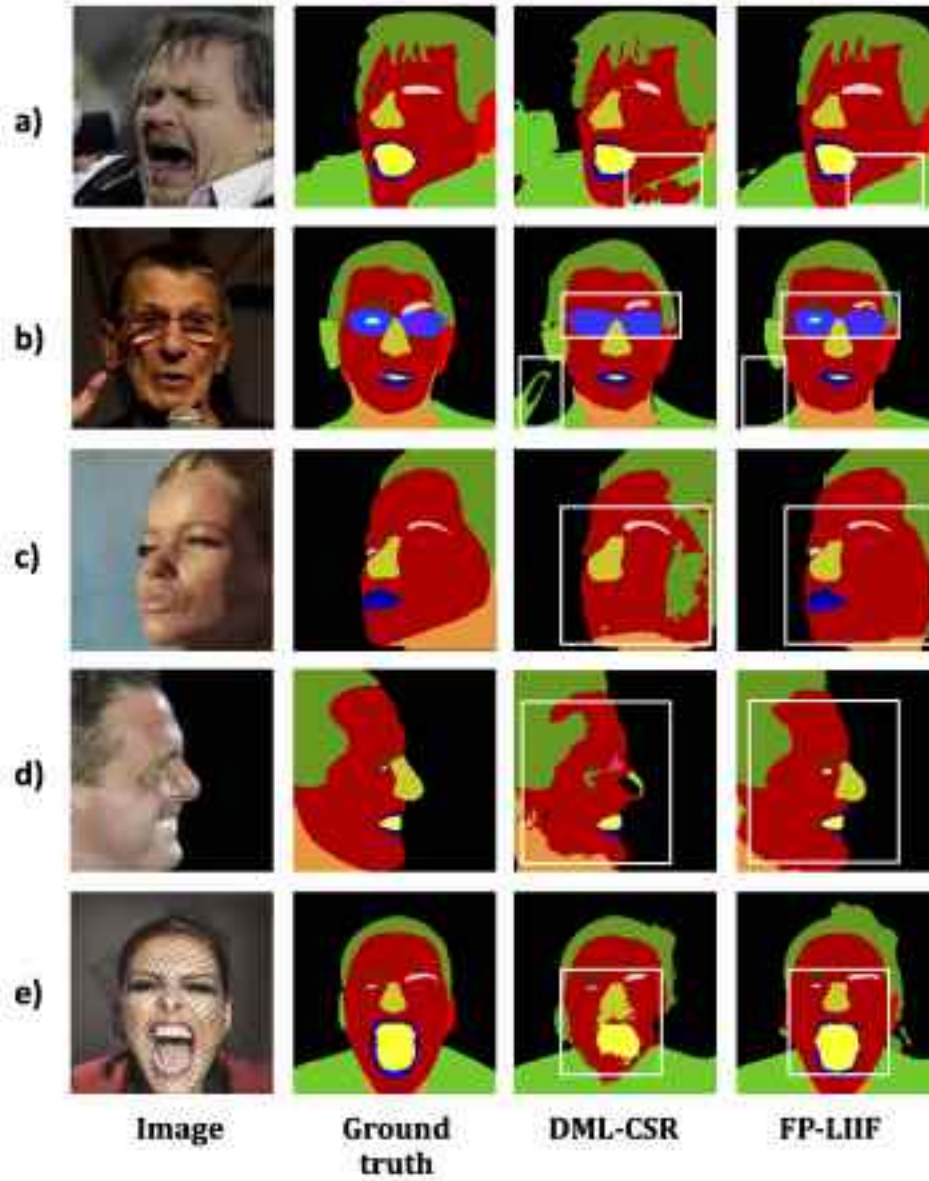

Figure 3. Visualization of a few results in CelebAMask-HQ dataset. The difference between DML_CSR and our results is highlighted. The cloth region in a), b), eyes in b),c) d) and nose in d),e) are better predicted by FP-LIIF
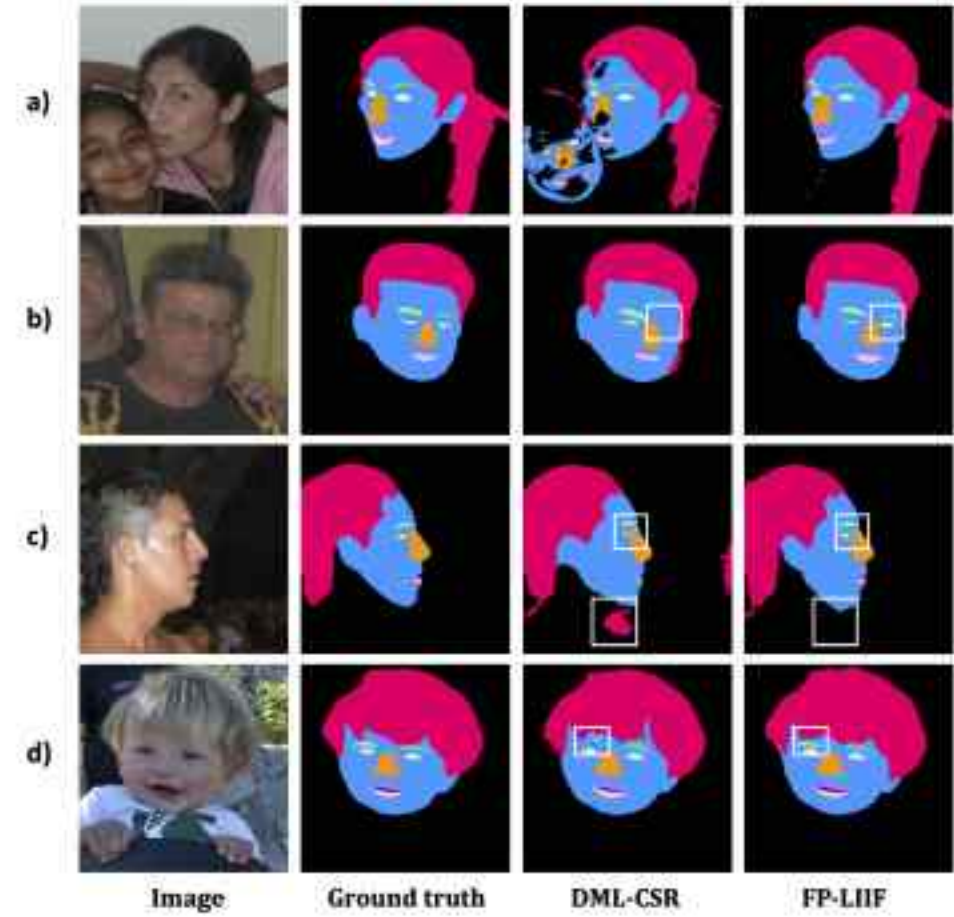

Figure 4. Visualization of a few results in LaPa dataset. The difference between DML_CSR and our results is highlighted. Eyes in b),c),d), Brows in b),d) are better predicted by FP-LIIF

simultaneously, some approaches model the individual classes separately. These approaches, called the local methods, claim that focusing on the facial components (e.g. eyes, nose, etc.) results in more accurate predictions but at the expense of efficient network structure in terms of parameter sharing. Luo et al. [32] propose a model which segments each detected facial part hierarchically into components and pixel-wise labels. Zhou et al. [60] built interlinking CNNs to perform localization followed by labeling. Lin et al. [25] propose an RoI-Tanh operator-

based CNN that efficiently uses backbone sharing and joint optimization to perform component-wise label prediction. Our proposed method is a whole image-based method that uses a single encoder-decoder pair to parse faces using implicit neural representations on the global image scale.

2.2. Parametric Human Face Models and Implicit representations

Parametric models for the human face have been explored for a long time since the pioneering work of [39]. Principle component analysis (PCA) was used to model face geometry and appearance in [1] to create a 3D Morphable model (3DMM) of the face. Many of the approaches where the 3D face model is estimated from the 2D image, estimate coefficients of the pre-computed statistical face models [41,50,51]. Other methods use regression over vox-
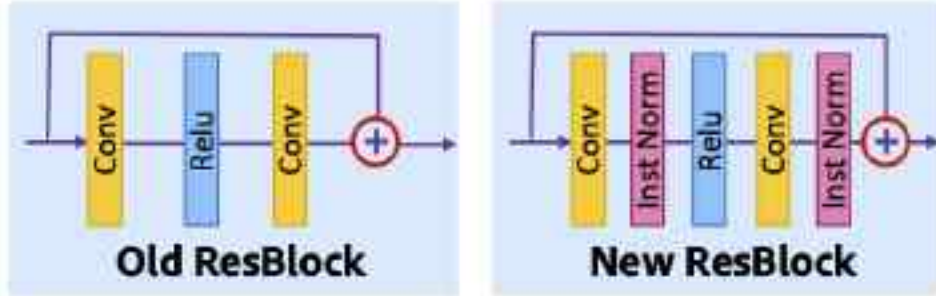
Figure 5. ResBlock Modification: Comparison of the residual block design in EDSR with our modification. We add an Instance normalization after each convolution in the residual block.

els or 3D meshes [10, 53] of face to arrive at a detailed face model. Many approaches have also used 3DMM type models with deep learning techniques [6, 7, 9, 18, 45].

With the emergence of Implicit Neural Representation [33, 34, 38, 47], a new approach to parameterizing signals has been gaining popularity. Instead of encoding signals as discrete pixels, voxels, meshes, or point clouds, implicit neural representation can parameterize these as continuous functions. A lot of work in the field has been around 3D reconstruction and shape modeling [4, 15, 30, 33, 34, 38, 55]. Deepsdf [38] encodes shapes as signed distance fields and models 3D shapes with an order of magnitude lesser parameters. Neural Volume [30] and NeRF [34] introduced 3D volume rendering by learning a continuous function over 3D coordinates. These works led to the use of implicit representation in the domain of human body or face rendering like [43, 56], that use implicit representation to model human heads and torso in a detailed manner. Others like Pi-Gan [2] and NerFACE [11] used it in the domain of faces. NerFACE [11] can extract a dynamic neural radiance field face from a monocular face video and be used with the parameters of a 3DMM. Besides 3D modeling, implicit neural representation has also been used in 2D image-to-image translation. Local Implicit Image function (LIIF) [3] proposed an implicit neural representation-based super-resolution model that treats images in the continuous domain. Based on these approaches of low-dimensional parametric face models, stunning performance of implicit neural representation in 3D reconstruction, and 2D image-to-image translation, our choice of method for exploring face segmentation gravitated towards the implicit representation approach of LIIF. The 2D texture-less appearance of the face segmentation mask prompted us to explore a low-parameter version of the LIIF model for face parsing.

3. Methodology

The human face has a regular and locally consistent structure because the various features on the human face, like eyes, nose, mouth, .etc, would maintain their relative position. We use this uniformity to design a lightweight model for face parsing. We adopt the LIIF [3] framework to

learn a continuous representation for segmentation for locally consistent structures of human faces.

3.1. Segmentation as Local Implicit Image Function

An image I in LIIF is represented by a 2D grid of features $Z \in \mathbb{R}^{H \times W \times D}$ such that a function f_θ can map each $z \in \mathbb{R}^D$ in Z to another domain. Here, f is an MLP, and θ are its parameters. This can be represented by eq 1:

$$s = f_\theta(z, x) \tag{1}$$

where, $x \in \mathcal{X}$ is a 2D coordinate, and s is the signal in the domain we want to convert our image I into. The coordinates are normalized in the $[-1, 1]$ range for each spatial dimension. In this paper, s is the probability distribution among a set of labels, i.e., $P(y|I, x)$, where y denotes the class label. So for a given image I, with latent codes z and query coordinate x_q, the output can be defined as $P(y|x_q) = f_\theta(z, x_q)$. So using the LIIF approach, we can write

$$P(y|x_q) = f_\theta(z^*, x_q - v^*) \tag{2}$$

where z^* is the nearest z to the query coordinate x_q and v^* is the nearest latent vector coordinate.

Other methods mentioned in LIIF [3], such as Feature Unfolding and Local Ensemble, are also used. Feature Unfolding is a common practice of gathering local information or context by concatenating the local z in the 3×3 neighborhood for each z in Z. To illustrate, the feature unfolding of a Z of dimension $(H \times W \times D)$ would end up as $(H \times W \times 9D)$. Local Ensemble is a way to address the discontinuity in f_θ along sharp boundaries. An average of f_θ is calculated for each pixel according to the four nearest neighbors of z^*. This also bakes in a voting mechanism in the model at a per-pixel level.

3.2. Image Encoder

We now describe our image encoder that takes as input an RGB image of size 256×256 and generates an output volume of latent vectors of size 64×64. Our encoder is a modified version of EDSR [24] as shown in Figure 2. We modify all the resblocks by appending an instance normalization block [52] after every convolution layer, Figure 5. We create 24 resblocks Figure 2, and all convs have a size of 3×3 and filter depth of 64 channels unless otherwise stated. The input is first passed through a conv before passing it into resblock-groups.

We have three resblock-groups. We added the first two to extract and preserve the fine-grained information from the image while the activation volume undergoes a reduction in the spatial dimensions because of the strides conv. The third group of resblock is used to generate the image representation Z. Each of the resblock-groups are a series of resblocks followed by a residual connection from the input,
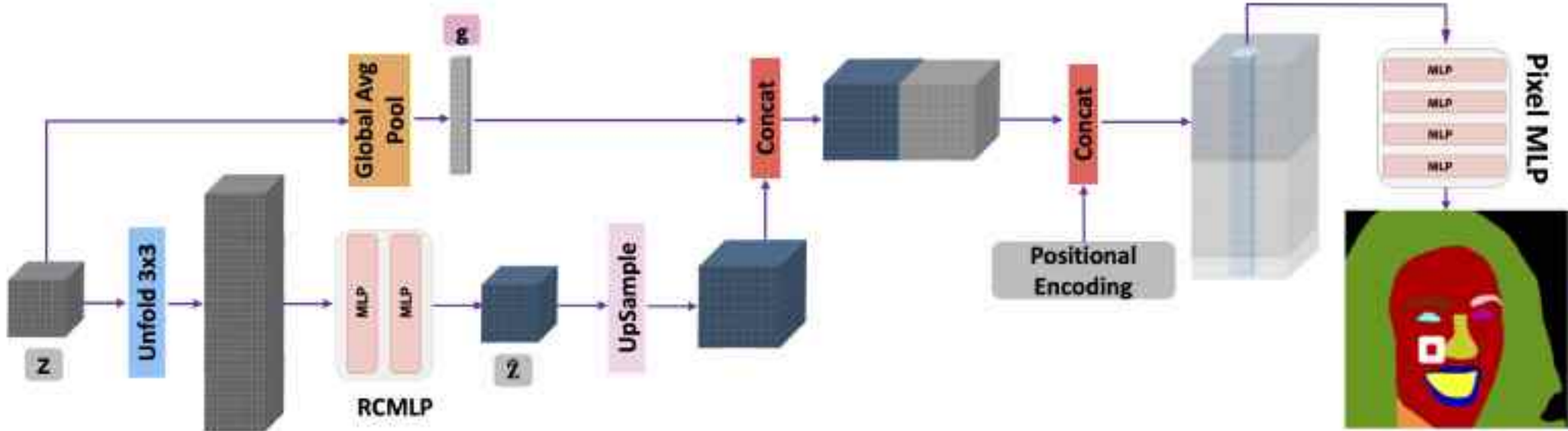
Figure 6. Decoder Architecture: Our decoder takes in the feature grid Z from the encoder and performs unfolding and global avg-pooling as shown in the figure. A two-layer fully connected MLP is used to reduce the number of channels in the unfolded volume. This is upsampled and concatenated with the global pool feature g and positional encoding. Finally, a four-layer MLP is applied per spatial location to generate the classwise probability distribution.
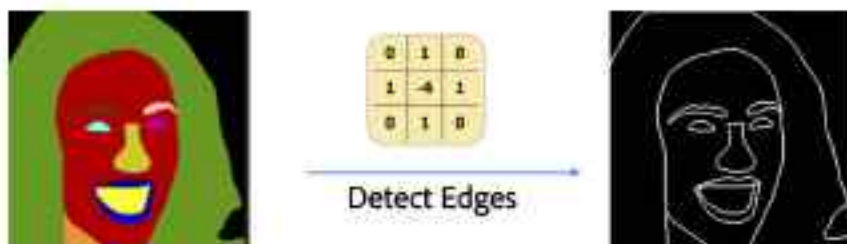

Figure 7. Binary edge generation

Figure 2. The output of the first resblock-group that contains two resblocks is passed to a 3×3 conv with a stride of 2. This is passed to the second resblock-group which has six resblocks. This is again followed by a conv of stride 2. The output of this second downsampling is passed through the third resblock-group containing 16 resblocks. This generates a feature volume of size $64 \times 64 \times 64$, which is then passed to the LIIF decoder.

3.3. LIIF decoder

The task of the decoder is to predict the segmentation labels at each pixel, which depends on the local context and the global context. Therefore, to provide global context during per-pixel prediction, we first extract the global context by doing an average pool of the latent volume along the spatial dimensions, as shown in Figure 6. Additional local context is added by passing the latent volume through a 3×3 unfolding operation, which increases the channel size to 64×9. The unfolded volume is then sent through a two-layer reduce channel MLP (RCMLP) with depths of 256 and 64. This makes the next upsampling operations computationally cheaper. The resulting volume $\hat{Z}$ of size $64 \times 64 \times 64$ is bilinearly upsampled to the output size and concatenated with the previously extracted global feature and two channels of positional encoding. The positional encodings are x,y coordinates ranging from -1 to 1 along the spatial dimension of the feature volume. This volume of

latent vectors is flattened and passed through a 4-layer MLP of 256 channels each to predict logits for the segmentation labels. Next, we perform a LIIF-like ensemble of these predictions using multiple grid sampling over $\hat{Z}$. Note that the regular conv can't be used to directly replace the unfolding operation because grid sampling result would differ for a $\hat{Z}$ derived from a $w \times h \times 9D$ volume compared to a $w \times h \times D$ volume because of the different neighbors of the D and $9D$ channels.

3.4. Loss

The logits are passed through a softmax and then guided with a cross-entropy loss L_{cce} and edge-aware cross-entropy loss L_{e_cce}. The edge-aware loss is calculated by extracting the edges of the ground truth label map with an edge detection kernel (Fig. 7) and calculating cross-entropy only on these edges. The final loss can be defined as:

$$L = L_{cce} + \lambda . L_{e\_cce} \qquad (3)$$

where λ is the additional weight for the edge-cross-entropy.

4. Experiments

4.1. Datasets

We use three face datasets to perform our experiments, LaPa [29], CelebAMask-HQ [21] and Helen [19]. LaPa is a face dataset with more in-the-wild photos having varying poses and occlusions. It has 22,168 images, of which 18,176 are for training, 2000 are for validation, and 2000 are for testing. The segmentation masks in LaPa have 11 categories: skin, hair, nose, left/right eyes, left/right brows, and upper/lower lips. The CelebAMask-HQ dataset contains 30k face images split into 24,183 training, 2993 validation, and 2824 test samples. It has a total of 19 semantic labels, including labels for accessories like eyeglasses,

Method/Class	Skin	Hair	Nose	I-Mouth	L-Eye	R-Eye	L-Brow	R-Brow	U-Lip	L-Lip	Mean
Wei et al. [54]	96.1	95.1	96.1	89.2	88.9	87.5	86	87.8	83.1	83.8	89.36
EAGR [49]	97.3	96.2	97.1	90	89.5	90	86.5	87	88.1	89	91.07
FARL^{Scratch} [59]	97.2	93.1	97.3	89.4	91.6	91.5	90.1	89.7	87.2	89.1	91.62
AGRNET [48]	**97.7**	**96.5**	97.3	**90.7**	91.6	91.1	89.9	90	**88.5**	**90.1**	92.34
DML_CSR [58]	97.6	96.4	**97.3**	90.5	91.8	91.5	90.4	90.4	88	89.9	92.38
Ours	97.6	96	97.2	90.3	**92**	**92.2**	**90.9**	**90.6**	87.8	89.5	**92.41**
Ours^{512}	97.5	95.9	97.2	90.3	**92**	**92.2**	**90.9**	**90.6**	87.7	89.5	**92.38**
Ours$^{192\to256}$	97.5	96	97.2	90.3	**92**	**92.1**	**90.8**	**90.5**	87.7	89.4	92.35
Ours$^{128\to256}$	97.5	96	97.2	90.2	91.6	**91.8**	**90.8**	**90.4**	87.5	89.3	92.23
Ours$^{96\to256}$	97.4	95.9	97.2	90.0	90.1	90.5	**90.5**	90.2	87	88.9	91.76
Ours$^{64\to256}$	97.1	95.8	97	89.4	85.5	86.2	88.8	88.6	85	87.8	90.12

Table 1. Results on Lapa: F1 score comparison with baselines. Ours^{512} denotes the result of generating output at 512 resolution without changing the input resolution of 256×256. Ours$^{192\to256}$,Ours$^{128\to256}$,Ours$^{96\to256}$ and Ours$^{64\to256}$ denote results of upsampling from output resolution 192,128,96 and 64 respectively to a resolution of 256.

| Method/Class | Skin | Nose | E-glasses | L-Eye | R-Eye | L-Brow | R-Brow | L-Ear | R-Ear | Mean |
	Mouth	U-Lip	L-Lip	Hair	Hat	Earring	Necklace	Neck	Cloth	
Wei et al. [54]	96.4	91.9	89.5	87.1	85	80.8	82.5	84.1	83.3	82.05
	90.6	87.9	91	91.1	83.9	65.4	17.8	88.1	80.6	
FARL^{Scratch} [59]	96.2	93.8	92.3	89	89	85.3	85.4	86.9	87.3	84.77
	91.7	88.1	90	94.9	82.7	63.1	33.5	90.8	85.9	
EAGR [49]	96.2	94	92.3	88.6	88.7	**85.7**	85.2	88	85.7	85.14
	95	88.9	91.2	94.9	87.6	68.3	27.6	89.4	85.3	
AGRNET [48]	96.5	93.9	91.8	88.7	89.1	85.5	85.6	88.1	**88.7**	85.53
	92	89.1	91.1	95.2	87.2	69.6	32.8	89.9	84.9	
DML_CSR [58]	95.7	93.9	**92.6**	89.4	89.6	85.5	**85.7**	**88.3**	88.2	**86.07**
	91.8	87.4	91	94.5	**88.5**	**71.4**	40.6	89.6	85.7	
Ours	**96.6**	**94**	92.4	**89.6**	**89.7**	85.2	84.9	86.7	86.6	**86.07**
	92.6	**89.1**	91.1	**95.2**	86.8	66.9	**43.9**	**91.3**	**86.7**	
Ours^{512}	**96.6**	**94**	92.5	**90**	**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**	
Ours$^{192\to256}$	**96.6**	**94**	92.4	**89.6**	**89.7**	85.2	84.9	86.7	86.6	86.05
	92.5	**89.1**	91.1	**95.2**	86.8	66.9	43.8	**91.3**	86.6	
Ours$^{128\to256}$	**96.6**	93.9	92.4	**89.6**	89.6	85.2	84.9	86.7	86.6	86.03
	92.5	88.9	91	**95.2**	86.8	66.9	**43.9**	**91.3**	86.6	
Ours$^{96\to256}$	**96.5**	93.9	92.3	89.3	89.3	85.1	84.8	86.7	86.5	85.90
	92.4	88.6	90.9	95.1	86.7	66.7	**43.6**	**91.3**	86.6	
Ours$^{64\to256}$	**96.4**	93.7	92.1	88.5	88.5	84.5	84.3	86.4	86.3	85.52
	92.1	87.5	90.3	95.1	86.6	65.7	**43.7**	**91.2**	**86.5**	

Table 2. Results on CelebAMask-HQ: F1 score comparison with baselines. Ours^{512} denotes the result of generating output at 512 resolution without changing the input resolution of 256×256. Ours$^{192\to256}$,Ours$^{128\to256}$,Ours$^{96\to256}$ and Ours$^{64\to256}$ denote results of upsampling from output resolution 192,128,96 and 64 respectively to a resolution of 256.

necklaces, earrings, etc., in addition to the labels in LaPa. Helen is the smallest of these three with 11 categories and has 2330 images with 2000, 230, and 100 in the training, validation, and test samples respectively.

4.2. Implementation Details

We implement our training and evaluation pipeline on PyTorch [40] version 1.10.0 with the CUDA 11.1 backbone

Model/Class	Skin	Nose	U-lip	I-mouth	Overall
	L-Lip	Eyes	Brows	Mouth	
EAGR	94.6	96.1	83.6	89.8	93.2
	91	90.2	84.9	95.5	
DML-CSR	96.6	95.5	87.6	91.2	93.8
	91.2	90.9	88.5	95.9	
Ours	95.1	94	79.7	86.3	91.2
	87.6	89.1	81	93.6	

Table 3. Results on Helen: F1 score comparison with baselines. Our results on non aligned Helen face data set are comparable to SOTA.

Model	Params ↓	×FP-LIIF ↓	GFlops ↓	FPS↑
DML_CSR	59.67 M	26	253	76
EAGR	66.72 M	29	235	71
FARL	150 M	65	370	26
FP-LIIF (ours)	**2.29 M**	**1**	**85**	**110**

Table 4. Model size comparison: The table shows the parameter count, GFlops, and FPS for each of the models and the relative size of each model compared to FP-LIIF
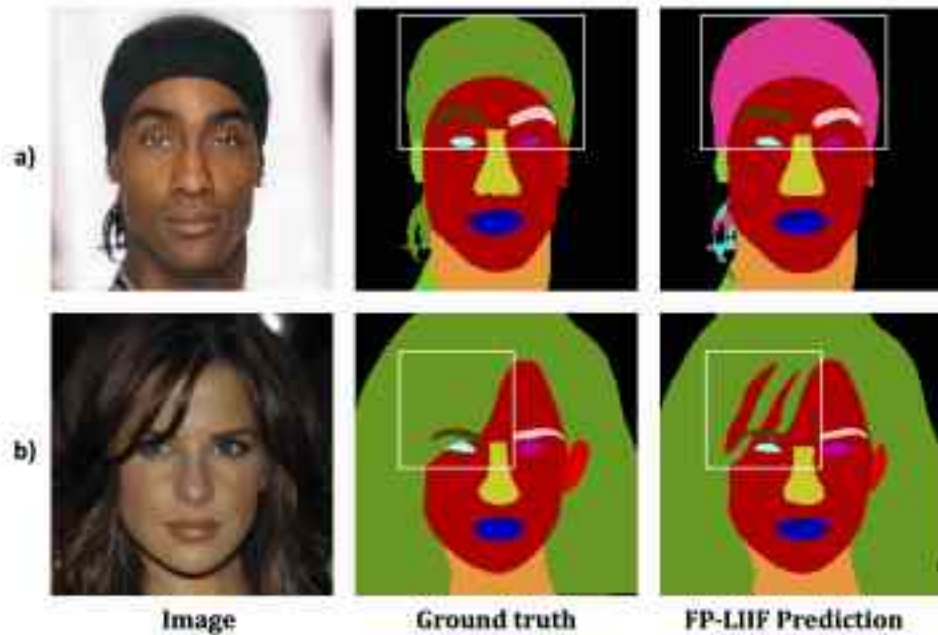

Figure 8. Few test samples from CelebAMask-HQ dataset illustrating noisy ground truth data and our prediction for the same. In the top row headgear has been marked as hair and in the bottom row strands of hair are not clearly segmented in the ground truth mask.

on Python 3.8.5. We train FP-LIIF on 4 Nvidia A-100 GPUs with a mini-batch size of 33 and 64 for CelebAMask-HQ and LaPa respectively. The network is optimized for 400 epochs using Adam [17] with an initial learning rate of 5e-4. The learning rate is decreased by a factor of 10 after every 20 epochs. The λ for edge-cross-entropy was set to 10 and 40 for CelebA and Lapa, respectively. Temperature scaling of softmax is also done with $\tau = 0.5$. The images and masks in the datasets were resized to 256×256 using bicubic sampling before being used for training and evaluation. During training various data augmentations are applied like random affine transformations of rotation by $30°$,

shear between 0 to 20, scaling between 0.5 to 3 followed by random cropping. Color jitter is also applied to the input image with a brightness between [0.5,1.5], contrast [0,2], saturation [0.5,1.5], and hue [-0.3,0.3].

4.3. Evaluation Metrics

To keep our evaluation consistent with other works, we primarily use a class-wise F1 score and a mean F1. In addition to that, we use mean intersection over union (mIoU) to compare with DML_CSR. The background class is ignored in all these metrics.

4.4. Baselines

We compare our FP-LIIF performance with several baselines, like Wei et al. [54] (figures taken from [59]), AGR-NET [48], EAGR [49], DML_CSR [58], FARL [59] from scratch, i.e., no pre-training. The results on LaPa are reported in Table 1, results comparing performance on CelebAMask-HQ are in Table 2 and results on Helen are in Table 3. Finally, a comparison of model size, Gflops and FPS is made in Table 4.

5. Results

According to Table 1's LaPa results, FP-LIIF performs better overall in mean-F1 and in classes such as eyes (left-eye and right-eye), brows (left-brow and right-brow), and skin. Table 2 demonstrates that our approach performs better than the baselines on CelebA in terms of mean-F1 and also at the class-level F1 of skin, nose, eyes (left-eye, right-eye), lips(upper-lips, lower-lips), hair, necklace, neck, and cloth. We have also included a row of results demonstrating our performance when we change our output size to 512×512. The results show that even without training for a higher resolution output, our network seamlessly generates decent segmentation results at a higher resolution with nominal degradation in LaPa while still matching the current SOTA of 92.38 by DML_CSR. Table 2 demonstrate superior performance at 512 resolution with a mean-F1 of 86.14, which is 0.07 higher than DML_CSR. We achieve these results without training on multiple resolutions, i.e., we train on just 256×256 and the network seamlessly scales to multiple resolutions. Our results on the Helen dataset, which has a small number of training samples (2000), are in Table 3. Our performance is close to SOTA despite training on non-aligned face images. Last but not least, in Table 4, we comprehensively compare the model sizes, GFlops and FPS of all our baselines. With only 2.29 million parameters, FP-LIIF is the most compact face-parsing network available; it is 65 times smaller than FARL and 26 times more compact than DML_CSR. Our fps, evaluated on an Nvidia A100 GPU, stood at 110 frames per second, whereas DML CSR's performance was at 76 frames per second, and EAGR and

AGR-Net demonstrated fps of 71 and 70, respectively. Additional comparative analysis of our results with DML_CSR is included in the supplementary.

5.1. Ablations

To evaluate the effect of several components we conduct the following ablations.

Network without LIIF Decoder: We replaced the LIIF decoder with a Conv U-Net type decoder. The total parameter count of this model is 9.31M params (3x FP-LIIF). The Mean F1 for this model on LaPa dataset is 84.9 compared to 92.4 for our model.

EDSR ResBlock vs BN/IN ResBlocks: The Old-EDSR ResBlock network produces an F1 of 92.3 on the LaPa dataset. New ResBlock + BN produces 92.32 and ResBlock + IN produces 92.4. The slight improvement prompted us to use IN.

Edge-aware Cross-entropy and λ: The table 5 indicates the effect of λ on the modulation of the edge-aware cross-entropy loss.

λ	0	10	20	30	40
F1 on LaPa	91.73	92.2	92.29	92.34	92.4

Table 5. Effect of edge-aware loss modulated by λ on networks performance

Comparison with lightweight segmentation model: Table 9 shows the results for face segmentation on LaPa using SFNet [22] which is a recent lightweight segmentation network for cityscapes.

Class	SFNet	Ours		Class	SFNet	Ours
Skin	94.75	97.6		R-Eye	76.12	92.2
Hair	87.27	96.0		L-Brow	76.98	90.90
Nose	98.71	972.		R-Brow	73.8	90.60
I-Mouth	78.86	90.30		U-Lip	97.28	87.8
L-Eye	79.55	92.00		L-Lip	96.23	89.5
				Mean	85.96	92.41

Figure 9. Comparison with SFNet [22], another lightweight segmentation network.

5.2. Low Resource Inference

One of the practical advantages of FP-LIIF is that it enables face parsing on low-resource devices. The primary prerequisite to enable low resource inference is that a model's inference should be low in compute and therefore have a high frame per second (FPS) count. Our ability to predict segmentation masks at multiple resolutions enables us to meet the demand for low inference costs. To achieve this, we can instruct the network to perform a low-resolution prediction, and the result can be upscaled to a higher resolution. Table 6 shows the FPS for lower-resolution inference

on a single Nvidia A100 GPU. However, the shorter inference time should not result in poor quality output when upsampled to a higher resolution. Therefore we compare our upscaled outputs with the ground truth and present the findings in Table 1, 2. Here, we can see that our 192×192 or 128×128 segmentation output, when upscaled to 256×256, leads to a minimal loss in quality, as can be seen in both classwise and overall F1 scores. In Table 1, 2 Ours$^{192 \rightarrow 256}$ denote results of upsampling from 192×192 to 256×256 and similarly Ours$^{128 \rightarrow 256}$, Ours$^{96 \rightarrow 256}$ and Ours$^{64 \rightarrow 256}$ denote results of upsampling from 128, 96 and 64 respectively to 256×256. In addition to faster inference, a low parameter count or smaller size model helps in model transmission under low bandwidth circumstances.

Res	64x64	96x96	128x128	192x192	256x256
FPS	445	332	294	187	110
FLOPS	27.44	32.25	39	58.24	85.2

Table 6. Frame Per Second for different resolution output while keeping the input image resolution constant at 256×256

5.3. Limitation

Encouraged by the performance of this low parameter FP-LIIF network in face segmentation, we tested its effectiveness at semantic segmentation in a more generic domain like Cityscapes [5]. We chose this dataset because it lacked the structural regularity that we exploited in this work and segmentation using the current architecture should not be feasible. As expected, the mIoU score on the validation was reported at 62.2, which is 20+ points lower than SOTA models reporting scores in the range of ∼85 mIoU.

6. Conclusion and Future Work

This work presents FP-LIIF, an implicit neural representation-based face parsing network. We exploit the human face's regular and locally consistent structure to propose a low-parameter, local implicit image function-based network that can predict per-pixel labels for a human face image. Our model is $1/26^{th}$ or lesser in size compared to the current SOTA models of face parsing but outperforms them on mean-F1 and mean-IoU over multiple datasets like CelebAMask-HQ and LaPa. This network can also generate outputs at multiple resolutions, which can be very useful in reducing the inference time of segmentation by generating output at a lower resolution. These properties make it feasible to use this architecture on low-resource devices.

Future work will address misprediction in regions with fewer class labels. We would also extend Implicit neural representation-based segmentation to domains with a regular and consistent structure, like medical imaging, human body parts, etc., and to domains where structure uniformity is not guaranteed, like in the wild images.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2, 3

[2] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2, 4

[3] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2, 4

[4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 4

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 4

[7] Abdallah Dib, Cédric Thébault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12819–12829, 2021. 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[9] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 4

[10] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018. 4

[11] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2, 4

[12] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2

[13] Tianchu Guo, Youngsung Kim, Hui Zhang, Deheng Qian, ByungIn Yoo, Jingtao Xu, Dongqing Zou, Jae-Joon Han, and Changkyu Choi. Residual encoder decoder network and adaptive prior for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1

[14] Aaron S Jackson, Michel Valstar, and Georgios Tzimiropoulos. A cnn cascade for landmark guided semantic part segmentation. In *European Conference on Computer Vision*, pages 143–155. Springer, 2016. 1

[15] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 4

[16] Ira Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Transactions on Graphics (TOG)*, 35(4), 2016. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[18] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020. 4

[19] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *ECCV*. Springer, 2012. 5

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 1

[21] 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. 2, 5

[22] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *ECCV*, 2020. 8

[23] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3911–3919, 2017. 1

[24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 4

[25] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5654–5663, 2019. 1, 3

[26] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild, 2021. 1

[27] Sifei Liu, Jianping Shi, Ji Liang, and Ming Hsuan Yang. Face parsing via recurrent propagation. In British Machine Vision Conference 2017, BMVC 2017, British Machine Vision Conference 2017, BMVC 2017. BMVA Press, 2017. 1

[28] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3451–3459, 2015. 1

[29] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In AAAI, pages 11637–11644, 2020. 1, 2, 5

[30] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751, 2019. 4

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. 1

[32] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 2480–2487, 2012. 3

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4460–4470, 2019. 4

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020. 2, 4

[35] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF international conference on computer vision, pages 7184–7193, 2019. 1

[36] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 98–105. IEEE, 2018. 1

[37] Xinyu Ou, Si Liu, Xiaochun Cao, and Hefei Ling. Beauty emakeup: A deep makeup transfer system. In Proceedings of the 24th ACM International Conference on Multimedia, page 701–702, New York, NY, USA, 2016. Association for Computing Machinery. 1

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 165–174, 2019. 4

[39] Frederick Ira Parke. A parametric model for human faces. Technical report, UTAH UNIV SALT LAKE CITY DEPT OF COMPUTER SCIENCE, 1974. 3

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 6

[41] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. IEEE transactions on pattern analysis and machine intelligence, 43(11):4142–4160, 2020. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021. 2

[43] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5620–5629, 2021. 2, 4

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015. 1

[45] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multiview geometry consistency. In European Conference on Computer Vision, pages 53–70. Springer, 2020. 4

[46] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Conference on. IEEE, 2017. 1

[47] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In Proc. NeurIPS, 2020. 2, 4

[48] Gusi Te, Wei Hu, Yinglu Liu, Hailin Shi, and Tao Mei. Agrnet: Adaptive graph representation learning and reasoning for face parsing. IEEE Transactions on Image Processing, 30:8236–8250, 2021. 1, 6, 7

[49] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In European Conference on Computer Vision, pages 258–274. Springer, 2020. 1, 2, 6, 7

[50] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time

face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[51] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*, 1(2), 2019. 3

[52] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 4

[53] Huawei Wei, Shuang Liang, and Yichen Wei. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562*, 2019. 4

[54] Zhen Wei, Si Liu, Yao Sun, and Hefei Ling. Accurate facial image parsing at real-time speed. *IEEE Transactions on Image Processing*, 28(9):4659–4670, 2019. 6, 7

[55] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 4

[56] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 2, 4

[57] He Zhang, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, 127(6):845–862, 2019. 1

[58] Qi Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Computer Vision and Pattern Recognition*, 2022. 1, 2, 6, 7

[59] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 1, 2, 6, 7

[60] Yisu Zhou, Xiaolin Hu, and Bo Zhang. Interlinked convolutional neural networks for face parsing. In *Advances in Neural Networks – ISNN 2015*, pages 222–231. Springer International Publishing, 2015. 3

[61] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer graphics forum*, volume 37, pages 523–550. Wiley Online Library, 2018. 2
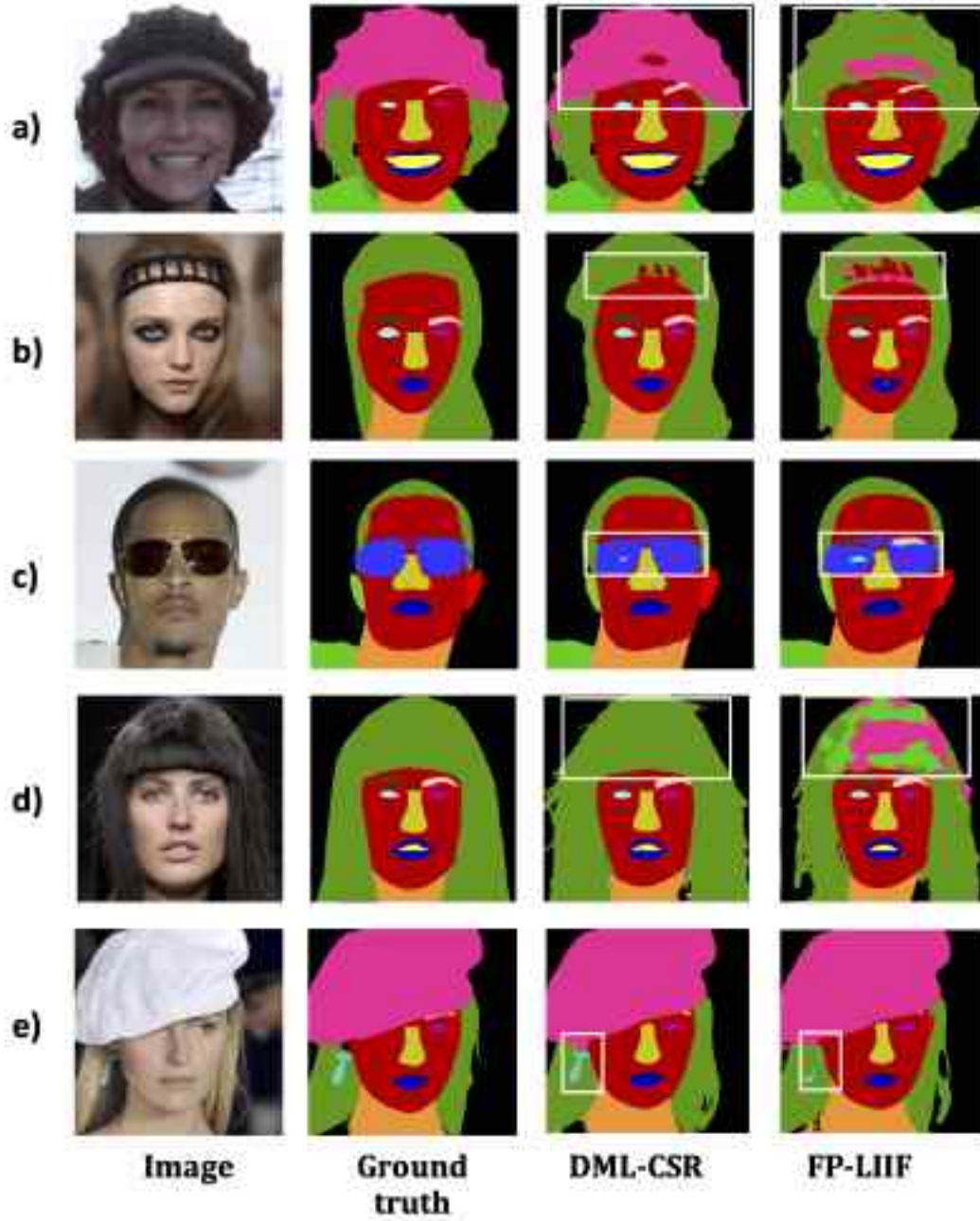
6.1. Analysis

Figure 10. Few results where DML_CSR performed better than FP-LIIF on CelebAMask-HQ dataset.

The quantitative results shown in Table 1, 2 points that even though FP-LIIF fares better in mean F1, the best classwise performance is scattered across multiple models. But at the same time, the gap between the best classwise scores and FP-LIIF's classwise scores is marginal. Therefore, we try to further identify the problematic areas and include visualizations of FP-LIIF's worst-performing results compared to DML_CSR in F1 in Figure 11, 10. It can be seen from Figure 11 that rows a) and c) have negligible differences, and in the remaining rows, both are performing poorly in the problematic regions of hair and face. In Figure 10's rows b) and d), the F1 scores for these are debatable because of incorrect labeling in the ground truth. In the remaining rows, the underrepresented class of hat and earrings are bringing down our performance. Therefore the current setup of FP-LIIF is affected by a lack of data as compared to DML_CSR. This can also be corroborated by Table 2. It is also necessary to point out that the ground truth data of CelebAMask-HQ is noisy (Figure 8) and can cause problems in training and testing.

From the point of view of inference time, FP-LIIF could be used to generate segmentation at a lower resolution, and
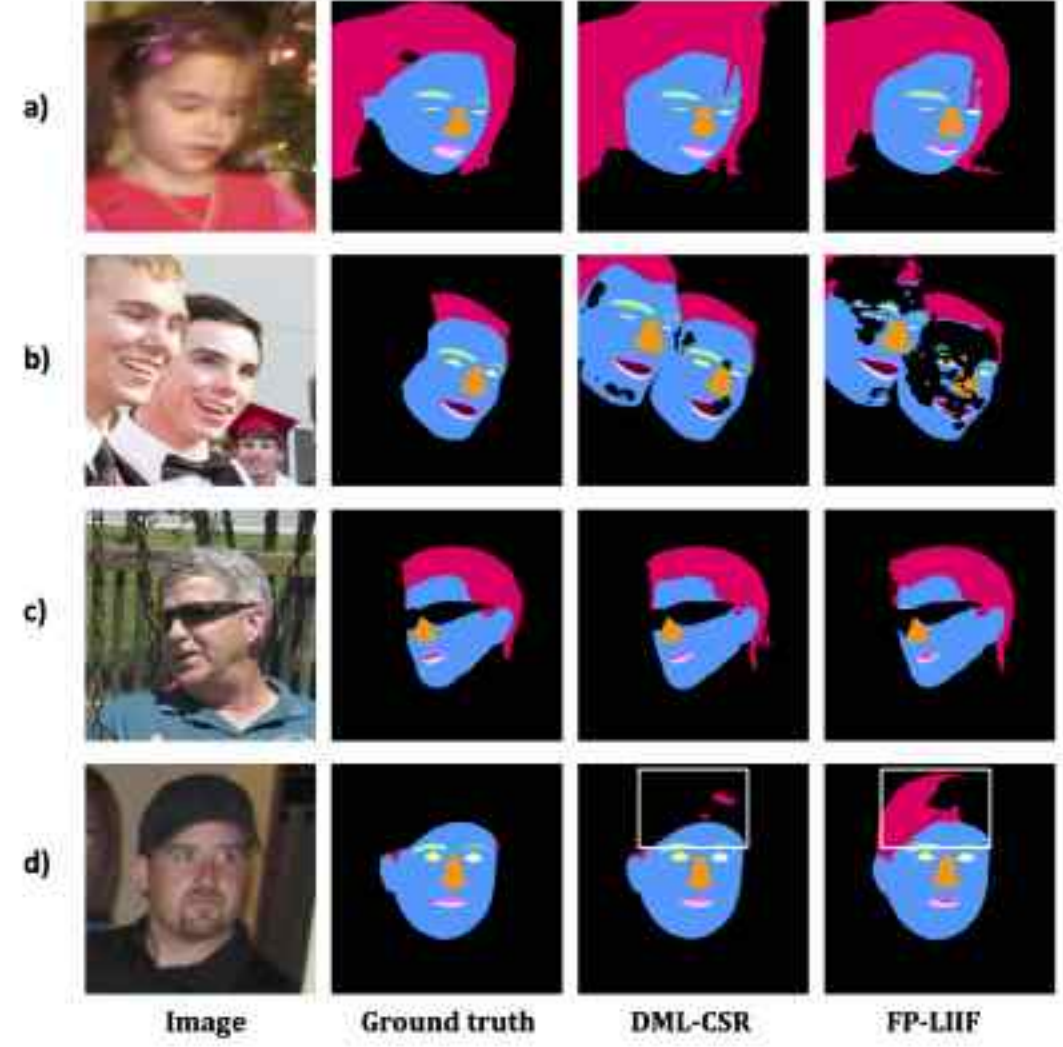

Figure 11. Few results where DML_CSR performed better than FP-LIIF on LaPa dataset

the generated output scaled at the required higher resolution to improve inference time and hence increase fps. The generation of lower-resolution segmentation does not require any additional training and is an outcome of being an implicit neural representation network. The 128-resolution version of FP-LIIF clocked an fps of 294 compared to the regular version of resolution 256, which runs at 120 fps. This makes our model more conducive for low compute devices.

6.1.1 Variance in performance over mulplte runs

We also calculate the mean and variance of our model's F1 score for Lapa, CelebAMask-HQ and Helen in Table 7 It

	Mean	SD
F1 Lapa	92.35	0.06
F1 Celeb	85.90	0.20
F1 Helen	91.12	0.10

Table 7. Mean and Variance of FP-LIIF

should be noted that other state-of-the-art works do not report these mean and variance over multiple runs and therefore direct comparison of these numbers is not possible.