EasyPortrait: Face Parsing and Portrait Segmentation Dataset

Karina Kvanchiani, Elizaveta Petrova, Karen Efremyan, Alexander Sautin and Alexander Kapitanov
SberDevices, Moscow, Russian Federation

Keywords: Portrait Segmentation, Face Parsing, Portrait Dataset.

Abstract: Video conferencing apps have recently improved functionality by incorporating computer vision-based features such as real-time background removal and face beautification. The lack of diversity in existing portrait segmentation and face parsing datasets – particularly regarding head poses, ethnicity, scenes, and video conferencing-specific occlusions – motivated us to develop a new dataset, EasyPortrait, designed to address these tasks simultaneously. It contains 40,000 primarily indoor photos simulating video meeting scenarios, featuring 13,705 unique users and fine-grained segmentation masks divided into 9 classes. Since annotation masks from other datasets were unsuitable for our task, we revised the annotation guidelines, enabling EasyPortrait to handle cases like teeth whitening and skin smoothing. This paper also introduces a pipeline for data mining and high-quality mask annotation through crowdsourcing. The ablation study demonstrated the critical role of data quantity and head pose diversity in EasyPortrait. The cross-dataset evaluation experiments confirmed the best domain generalization ability among portrait segmentation datasets. The proposed dataset and trained models are publicly available*.

1 INTRODUCTION

Video conferencing apps have quickly gained popularity in corporations for online meetings (Sander, 2020) and in daily life to keep in touch with distant relatives[1]. The video conferencing market value worldwide is expected to continue growing in the coming decades[2]. As a result, these apps have been enhanced with various beneficial features, including face beautification, background blur, and noise reduction[3]. Such extensions can improve the user experience by enabling background changing and skin smoothing[4].

Our work focused on incorporating the described features into the video conferencing app. The app should ensure a real-time CPU-based experience on the user's device and produce a highly accurate response. Besides, our system must be robust to the amount of context in images, backgrounds, persons in the frame, their poses, attributes (e.g., race and age), and accessories (headphones, hats, etc.). Finally, improving the users' experience with such functions as teeth whitening and accurate background changing in the case of transparent glasses lenses and uneven hair is preferable.

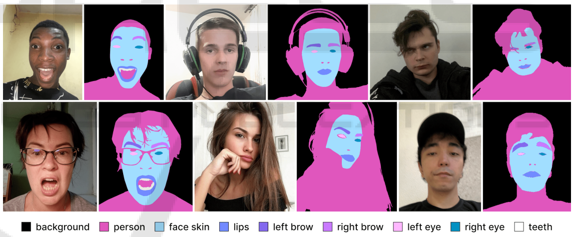

Figure 1: The face parsing and portrait segmentation annotation examples from the EasyPortrait dataset.

All these requirements impose restrictions on the solution and training dataset. The system must function in real-time without delays and produce fine-grained segmentation masks. The suitable data is required to be 1) heterogeneous in subjects, their head turns, subject-to-camera distances, scenes, and specific for videoconferencing domain subjects' accessories like eyeglasses and headphones; 2) annotated with main face parsing classes ("skin", "brows", "eyes", and "lips") and an extra class "teeth"; 3) accu-

*https://anonymous.4open.science/r/easyportrait-5109/

[1] https://inlnk.ru/Vo8ydV

[2] https://www.statista.com/statistics/1293045/video-conferencing-market-value-worldwide/

[3] https://www.banuba.com/blog/10-video-conferencing-trends-enhanced-by-face-ar-technology

[4] https://www.businessinsider.com/zoom-video-conferencing-filter-touch-up-work-from-home-coronavirus-2020-3

Kvanchiani, K., Petrova, E., Efremyan, K., Sautin, A. and Kapitanov, A.
EasyPortrait: Face Parsing and Portrait Segmentation Dataset.
DOI: 10.5220/0013135300003912
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 327-338
ISBN: 978-989-758-728-3; ISSN: 2184-4321

rately annotated for both tasks simultaneously. This approach enables training a single model to handle all potential use cases, achieving competitive metrics while saving limited resources for model inference (see Table 4 in the supplementary details).

Existing datasets are unsuitable for our purpose due to the limitations described in Section 2, which motivated us to create the EasyPortrait dataset. Images were annotated manually by 9 classes according to specially designed rules, which allowed us to cover all described cases. We checked that such data characteristics as the number of samples and their diversity in head pose positively impact the model's robustness and effectiveness (Section 4). The generalization ability of the training data was also assessed through cross-dataset evaluation experiments (Section 5).

Our contributions can be summarized as follows:

- We present EasyPortrait, a portrait segmentation and face parsing dataset containing 40,000 pairs of images and segmentation masks from 13,705 individuals in domain-suitable scenes with different head poses and various specific videoconferencing app accessories. We considered the importance of ethnic diversity in solving problems based on persons and their facial attributes and collected images from users of various countries and continents.
- We developed a pipeline for gathering and labeling images with fine-grained segmentation masks via crowdsourcing platforms. This approach balances annotation cost and quality, efficiently producing detailed and reliable image masks.
- The experiments demonstrate that EasyPortrait exhibits the best generalization capability regarding mIoU metrics across all portrait segmentation test sets.

2 RELATED WORK

This section discusses existing portrait segmentation and face parsing datasets separately. We will consider them from two points of view: 1) the applicability to the videoconferencing domain and 2) data creation techniques for each segmentation task and their consequences. We focus solely on image datasets, analyzing and benchmarking segmentation models on static images to maintain a consistent evaluation environment while avoiding video data's added complexity and computational demands.

2.1 Portrait Segmentation and Matting Datasets

The portrait segmentation task involves labeling every pixel in an image as either "person" or "background". Since matting annotations can be reduced to binary ones, image matting datasets will also be considered. As videoconferencing apps always take portraits for input and not a person entirely, only datasets with waist-deep people are reviewed. Therefore, other popular image person segmentation and image matting datasets, e.g., P3M-10k (Li et al., 2021), PPR10K (Liang et al., 2021), PhotoMatte13K/85 (Lin et al., 2020), Persons Labeled (per, 2020), are not described in this paper. We selected EG1800 (Shen et al., 2016), AiSeg (ais, 2019), FVS (Kuang and Tie, 2021), Face Synthetics (Wood et al., 2021), and HumanSeg14K (Chu et al., 2021) as the most widespread and predominantly containing photos of people from the waist up. Table 1 provides the numerical analysis of reviewed datasets.

Content. Chosen datasets can be divided into three groups regarding image source: 1) downloaded from websites, 2) collected manually, and 3) generated. The last two allow the data authors to determine content on their own directly. Images of EG1800 and AiSeg were collected from services like Flickr, which made their scenes multi-domain. The Face Synthetics dataset was entirely generated, entailing primarily blurred backgrounds, and thus is far from in the wild. The manually collected FVS (Kuang and Tie, 2021) and PP-HumanSeg14 (Chu et al., 2021) portrait segmentation datasets are single-domain with a bias towards videoconferencing. The FVS dataset provides composite images from 10 conference-style green-screen videos and virtual backgrounds. As a result, FVS samples suffer from the remaining green screen around a person in the frame. The PP-HumanSeg14 dataset includes 23 different conference backgrounds and samples of participants performing actions such as waving hands and drinking water. The provided samples contain one or more labeled people with faces blurred for privacy.

Annotation. Since segmentation mask annotation is one of the most challenging problems in the computer vision field, data authors prefer automatic methods. All reviewed datasets except PP-HumanSeg14K were annotated automatically or using Photoshop (see Fig. 5 in the supplementary material for visual examples). Such methods frequently produce coarse masks that prevent accurate high-frequency parts segmentation (e.g., hair) – one of the main hardships of background removal in video conferencing.

Table 1: Comparison of portrait segmentation and face parsing datasets. Because of the specifics of the tasks, we indicated the task name for the first ones and the number of classes for the second. Several datasets include images of different sizes, so the standard label was provided in the resolution column. 96% of images in the EasyPortrait are FullHD; see more information in Section 3.2. We also included notes about the annotation method, which can be important regarding label quality; more detailed information can be found in Section 2. Note that 20 classes in the FaceOcc dataset contain 19 classes from CelebAMask-HQ.

Dataset	Samples	Task / Classes	Resolution	Annotation Method
EG1800, 2016 (Shen et al., 2016)	1,800	segmentation	600 × 800	Photoshop
FVS, 2018 (Kuang and Tie, 2021)	3,600	segmentation	640 × 360	chroma-key
AISeg, 2018 (ais, 2019)	34,427	matting	600 × 800	automatically
PP-HumanSeg14K, 2021 (Chu et al., 2021)	14,117	segmentation	1280 × 720	manually
The Face Synthetics, 2021 (Wood et al., 2021)	100,000	segmentation	512 × 512	automatically
Helen, 2012 (Le et al., 2012)	2,330	11	400 × 400	automatically
LFW-PL, 2013 (Kae et al., 2013)	2,927	3	250 × 250	automatically & refined
CelebAMask-HQ, 2019 (Lee et al., 2019)	30,000	19	512 × 512	manually & checked & refined
LaPa, 2020 (Liu et al., 2020)	22,176	11	LR	automatically & refined
iBugMask, 2021 (Lin et al., 2021)	22,866	11	HR	manually
The Face Synthetics, 2021 (Wood et al., 2021)	100,000	19	512 × 512	automatically
FaceOcc, 2022 (Yin and Chen, 2022)	30,000	20	512 × 512	manually
EasyPortrait, 2023	40,000	9	FullHD	manually & checked

2.2 Face Parsing Datasets

The face parsing task aims to assign pixel-level semantic labels for facial images. Generally, face parsing refers to classifying image pixels, such as brows, eyes, nose, lips, mouth, and skin. We considered several widespread face parsing datasets for our purposes (Table 1 and Fig. 6 in the supplementary material).

Content. The main limitation of existing face parsing datasets is the low diversity in head poses and the absence of specific occlusions. Helen's (Le et al., 2012) authors obtained the data from other datasets and websites like Flickr by searching for "portrait" in various languages to avoid cultural bias. The CelebAMask-HQ dataset (Lee et al., 2019) mainly contains front-facing images of celebrities from Celeba (Liu et al., 2015) with centered heads. Besides, the faces usually occupy a significant part of the image. Thus, background information is mainly discarded. The LaPa dataset (Liu et al., 2020) was designed based on images from the 300W-LP (Zhu et al., 2017) and Megaface (Taherkhani et al., 2018) datasets. Received faces were aligned and mostly cropped, with limited background information preserved. This image collection method was also utilized to create the iBugMask dataset (Lin et al., 2021) containing samples from Helen's training set. The iBugMask authors focused on head pose diversity and augmented images with a synthetic rotation algorithm from 3DFFA (Guo et al., 2018), which led to massive artifacts. The Face Synthetics dataset was specifically diversified during generation by various head poses, human identities, clothes, and such occlusions as eyeglasses and face masks.

Annotation. Since the face parsing annotation process is more challenging than portrait segmentation, the data is frequently marked manually or automatically with further refining. Additionally, existing datasets had annotations created with inappropriate rules, limiting their usability for our purposes. First, almost all reviewed datasets relate beard to skin class and nostrils to nose class. Second, some contain glasses as skin and others – annotate transparent glasses as non-transparent ones. Such factors made the skin enhancement task impossible due to further artifacts. Third, none of them contain separate annotations for teeth. Finally, there are other difficulties present:

- LFW-PL (Kae et al., 2013) is limited to only 3 classes (background, face, hair), which is unsuitable for solving our specific problems.
- Helen's (Le et al., 2012) 2,330 facial images were annotated by facial part landmarks utilizing Amazon Mechanical Truck, and then masks were generated automatically. The LaPa's (Liu et al., 2020) authors pointed out Helen's annotation errors.
- CelebAMask-HQ (Lee et al., 2019) ignored occlusions on its own, however, the authors (Yin and Chen, 2022) solved this problem with the dataset extension – FaceOcc. It contains images from CelebAMask-HQ and is annotated with only one class – occlusions (eyeglasses, tongue, makeup, and others). Regardless, FaceOcc includes a beard to the skin class. We are consider FaceOcc as CelebAMask-HQ with FaceOcc.
- The iBugMask (Lin et al., 2021) contains many persons with annotated masks for only one of them.

he mentioned datasets are unsuitable for our task due to the outlined issues and additional challenges, including the absence of FullHD images, a limited number of subjects, privacy concerns, and low-quality annotations. In addition to general shortcomings, other datasets lack video-conferencing domain-specific characteristics like task-specific occlusions and situations, various head poses, and domain context scenes. Motivated by the above and the growing need for a suitable dataset tailored to video conferencing applications, we developed a new dataset, EasyPortrait, which includes both face parsing and portrait segmentation annotations. We intentionally diversified the EasyPortrait by head poses, subjects, scenes, subjects' attributes such as ethnicity, and their occlusions (glasses, beards, piercing, etc.). It was annotated with all required classes for our applications, with specific rules for the skin class and occlusions in particular (Table 3 in the supplementary material).

3 EasyPortrait DATASET

This part provides our dataset creation pipeline overview, the dataset characteristics, and its splitting.

3.1 Image Collection & Labeling

Two crowdsourcing platforms, Toloka[5] and ABC Elementary[6], were chosen for all stages of dataset creation. Our pipeline consists of two main stages: (1) the image collection stage, which is followed by validation completely realized on Toloka, and (2) the mask creation stage, which was accomplished on both platforms. All crowd workers are ensured fair compensation, reflecting their contributions and efforts. At each stage, the responses of low-skilled workers were subjected to our quality control methods. A more detailed description is provided below.

Image Collection. The crowd workers' task was to take a selfie or a photo of themselves in front of the computer. As we aimed to design a diverse dataset in terms of occlusions, races, and head turns and make it suitable to solve teeth whitening problems, one of the further conditions periodically supplemented the task:

- Occlusions. The sent photo should contain one of such characteristics as hands in front of the face, glasses, the tongue out, headphones, or hats.
- Head turns. The head on the sent photo should be turned in any direction at various angles.
- Teeth whitening. Random workers were asked to send photos with open mouths.
- Ethnicity. We recognized the importance of having a diverse dataset of facial images and ensured the inclusion of individuals from various countries.

Note that all workers have signed a document stating their consent to the photo publication before starting the tasks.

Image Collection Quality Check. We foresaw the possible dishonesty of the crowd workers and checked all images for duplicates by image hash comparison. The image collection quality check also includes image validation, where images are reviewed for compliance with conditions such as the distinctness of the face, the head being entirely in the frame, and the clarity of the frame. The validation stage was available to crowd workers only after training and examination tasks. Different users checked each image 3 to 5 times for at least 80% confidence.

Image Labeling. The annotation of portrait segmentation usually has several ambiguous instances, such as occlusions in front of the person, hand-held items, hats, headphones, hair, and others. Face parsing masks are also unclear due to occlusions in front of the face, including tongue, hair, eyeglasses, beard, etc. The annotation rules directly affect the final segmentation masks and the model trained on them. Rules for annotating each class and processing occlusions for workers are given in Table 3 in the supplementary material.

The labeling stage was divided into parts to simplify the annotation process for the workers. All images received after the collection stage were gradually sent to the annotation of individual pairs of classes: person and background, skin and occlusions (which include such things as eyeglasses, beard, tongue out, and others), eyes and brows, lips and teeth. After labeling, we separated the overall masks of eyes and brows into left and right ones using heuristics. Fig. 2 visualizes the mask creation stages.

Image Labeling Quality Check. We required all workers to complete training tasks for each pair to enhance mask quality. We analyzed the quality of crowd workers' training by automatically comparing masks from professional data annotators.

We requested ABC Elementary's qualified workers to label each image with the subsequent verification by the platform's experts. Due to a lack of trust in the platform's experts, whom we did not manage, and the shortage of qualified annotators capable of providing sufficient overlap, we incorporated low-skilled annotators into the pipeline with an overlap of 5. Thus, each image was annotated by 5 crowd workers for subsequent averaging and getting the best result. Segmentation masks were created from polygons. We aggregated the same annotations to one segmentation mask (see bottom of the Fig. 2), checked IoU (In-

[5] https://toloka.yandex.ru/

[6] https://elementary.activebc.ru/

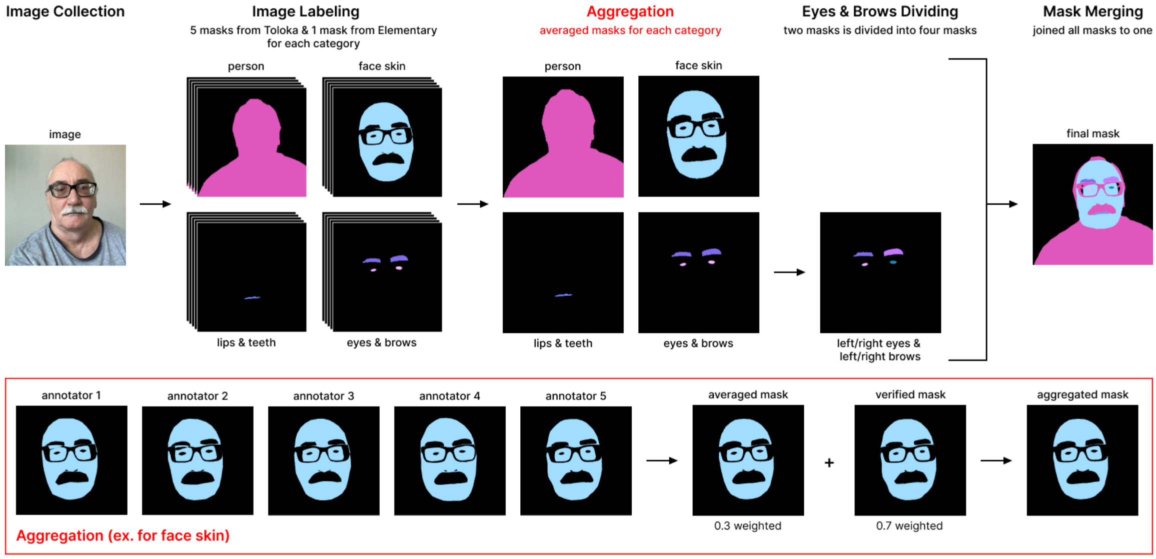

Figure 2: Example of data collection pipeline. Each image was annotated by 5 annotators. The masks are averaged with the expert-verified one and merged to obtain the final segmentation mask.

tersection over Union), and compared it to a unique threshold, selected for each pair manually[7]. Then, we averaged the received aggregated mask with verified one with weights of 0.3 and 0.7, respectively.

Mask Merging. Whole masks were received by the alternate overlay of masks in the following order: person, face skin, left brow, right brow, left eye, right eye, lips, and teeth.

In addition, the decision to release the dataset to the public and ethical reasons prompted us to add the filtration stage to the end of the pipeline – checking for children under 18, naked people, religious signs, and watermarks.

3.2 Dataset Characteristics

The mean and standard deviation of images in EasyPortrait are $[0.562, 0.521, 0.497]$ and $[0.236, 0.236, 0.232]$, respectively.

Classes. EasyPortrait is annotated with 9 classes, including “background”, “person”, “face skin”, “left brow”, “right brow”, “left eye”, “right eye”, “lips”, and “teeth”. We extracted all occlusions, such as glasses, hair, hands, etc., from the skin. The beard is extracted from the skin if it is clearly defined (refer to Fig. 7 in the supplementary material for details). However, such parts of a person as headphones, car belts, and others are included in the person class to facilitate background removal and streamline the data annotation process.

Diversity. The proposed dataset contains images of scenes such as an office, living room, kitchen, bedroom, outdoors, car, etc. Samples in the dataset display various clothes, hats, headphones, and medical masks). They are also diverse in lighting conditions, subjects’ age, gender, and poses. Most subjects in the dataset are between 20 and 40 years old, with approximately equal numbers of women and men (see Fig. 3h,k). Almost all images contain only one person, which is especially common at video conferences. We also collected images from regions such as Africa, Asia, India, and Europe, giving us approximate region and ethnic diversity (see Fig. 3a). Furthermore, some individuals in the photos display emotions, such as smiling, expressing anger, being surprised, and others (see Figure 3j). We examined the distribution of emotions using the facial emotional recognition model proposed by (Ryumina et al., 2022), which categorizes emotions into 7 classes: “neutral”, “happiness”, “sadness”, “anger”, “surprise”, “disgust”, and “fear”.

Images Resolution. We specifically focused on collecting FullHD images to avoid visual artifacts associated with rescaling. Most images in EasyPortrait, namely 38,353, are FullHD: the maximum side is 1,920, and the minimum side is 835 to 1,920 (Fig. 3e). The minimal resolution is 720 × 720. Fig. 3g shows the dataset’s samples’ mask area distribution.

Dataset Quality. We analyzed the number of points per image for each class since images were labeled with polygons. On average, each EasyPortrait image has 655 points, from which it can be concluded that the annotation is of high quality. In comparison,

[7] The thresholds were chosen by comparing crowd workers’ masks with corresponding experts’ masks from training tasks to ensure a qualitative visual result.

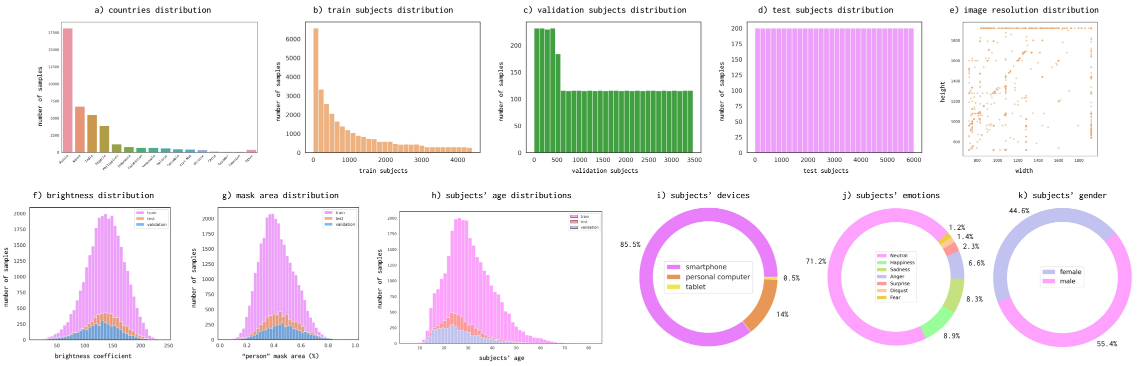

Figure 3: Dataset statistics. a) subjects' countries distribution; b), c), d) image distribution by subjects in train, validation, and test sets, respectively (each bar represents the count of images recorded by a particular subject group); e) image resolution distribution: samples overlap with equal transparency and density reveals quantity; f) brightness distribution; g) mask area distribution; h) subjects' age distribution; i) subjects' devices: only smartphones, personal computers, and tablets were used while recording; j) subjects' emotions; k) subjects' gender.

Helen (Le et al., 2012) was annotated only with 194 points per image and LaPa (Liu et al., 2020) with 106.

3.3 Dataset Splitting

All annotated images were divided into training, validation, and testing sets, with 30,000, 4,000, and 6,000 samples. Training images were received from 4,398 unique users, while validation and testing images were collected from 3,468 and 6,000 unique users, accordingly (see Fig. 3b-d). The test set contains the most unique users for maximum subject diversity. Additionally, the subjects in all three sets do not overlap, eliminating any possibility of data leakage. We also added the anonymized user ID hash, which can be used for manual dataset splitting.

4 ABLATION STUDY

We conducted an ablation study to assess the dataset's primary characteristics. By altering the dataset's data volume and head pose variability, we trained models and compared them with those trained on the original EasyPortrait. In the ablation study, we utilized BiSeNetv2 (Yu et al., 2018), FPN (Lin et al., 2017), FCN (Long et al., 2014), and Segformer-B0 (Xie et al., 2021) for portrait segmentation and face parsing tasks. Validation and test sets remain unchanged in all ablation experiments.

Quantitative Necessity. To evaluate the impact of data quantity, we trained selected models using varying training set sizes: 30,000 (original), 20,000, 10,000, and 5,000 images. The deterministic slice was used for a train set expansion, i.e., images in the $n[i]$ set are included in the $n[i+1]$ set. The ablation study results are provided in Fig. 4. The quantitative necessity experiments revealed an increase in metrics as the size of the training set expanded. Both portrait segmentation and face parsing metrics show an increase with the expansion of the training set size. However, the improvement in portrait segmentation is less prominent than the face parsing results.

Pose-Diversity Necessity. We also assess the importance of the head pose by varying pose diversity in 10,000 training images. We obtained the head pose coefficients (yaw and pitch) for each image in the dataset using 3DDFA network (Guo et al., 2018). For both of these coefficients, we chose three coefficient windows from homogeneous to heterogeneous pose distribution: [-7.5; 7.5], [-15; 15] [-180; 180]. Reducing head pose heterogeneity results in declining face parsing metrics (Fig. 4b). Variations in head rotations do not significantly impact portrait segmentation metrics; therefore, we did not include a plot for these experiments.

Cross-Dataset Ablation Study. We also conducted an additional set of experiments: we trained the FPN model on the EasyPortrait dataset with changes in data diversity. Then, we evaluated the model on other face parsing and portrait segmentation test sets, mentioned in Section 5.2. Alterations in data quantity and head pose variations have minimal impact on portrait segmentation results. In contrast, in the face parsing task, an increase in data diversity positively influences the model's metrics (Fig. 4a-b).

5 EXPERIMENTS

The main goal of extensive base experiments is to demonstrate that the dataset has the ability to train models, achieving concurrent results without the need to simulate facial occlusion or pose variations (as in

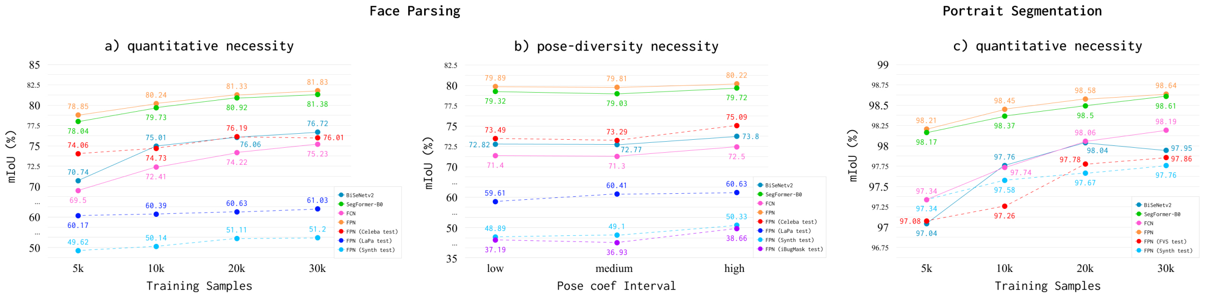

Figure 4: The impact visualization of such dataset characteristics as a) sample amount, b) head pose diversity for face parsing, and c) sample amount for portrait segmentation task. Solid lines correspond to models trained and tested on the EasyPortrait dataset. In contrast, the dotted line is the model pretrained on the EasyPortrait and tested on other datasets (see the legend for details). We evaluated all the datasets discussed in Section 5.2; however, we did not create visualizations for datasets without significant metric changes. Note that all the plots have different scales.

(Liu et al., 2020) and (Lin et al., 2021), respectively). For this reason, we chose various models for our base experiments. We evaluate the models' quality via the mean Intersection-over-Union (mIoU) metric (Long et al., 2015).

5.1 Base Experiments

Separation on Two Tracks. We split our experiments into two tracks – portrait segmentation and face parsing – to transparently compare EasyPortrait with other datasets separately. This division is also necessary to avoid ambiguity and ensure the obtained metrics represent both tasks. The portrait segmentation is based on two EasyPortrait classes ("background" and "person"), whereas the face parsing masks include eight classes ("background", "skin", "left brow", "right brow", "left eye", "right eye", "lips" and "teeth"). For portrait segmentation, we defined all classes of EasyPortrait except the background as a person, while for face parsing, we designated the person class as the background. The model configuration and training process are identical for both tasks, except for the number of classes in the decoder model's head.

Models. We prioritized lightweight architectures for easy integration into videoconferencing apps, enabling real-time use. As general segmentation architectures, we selected BiSeNetv2 (Yu et al., 2018), DeepLabv3 (Chen et al., 2017), FPN (Lin et al., 2017), FCN (Long et al., 2014), DANet (Fu et al., 2019), and Fast SCNN (Poudel et al., 2019) models. We utilized Segformer-B0 (Xie et al., 2021) to assess the performance of the transformer model on the proposed dataset. Besides the aforementioned widespread segmentation architectures, we experimented with models specifically designed for portrait segmentation and face parsing. For this purpose, we chose the SINet (Park et al., 2019a) and ExtremeC3Net (Park et al., 2019b) for the first one and EHANet (Luo et al., 2020) model for the second.

We trained each of these networks for 100 epochs with batch size 32. AdamW (Loshchilov and Hutter, 2017) was used as an optimizer and learning rate (LR) with the initial value of 0.0002. The LR changes according to the polynomial LR-scheduler with factor 1.0 by default. We also note that all random seeds in Python and PyTorch were fixed to a value of 1001 to enhance the reproducibility of the experiments.

Augmentations and Images Resolution. Images and segmentation masks were resized to the maximum side of 384 with aspect ratio preservation and symmetrically padded to square. We used bilinear interpolation for image resizing, while nearest neighbor interpolation was applied to masks to maintain class consistency. Photometric distortion was used with a brightness delta of 16, a contrast in the range [0.5, 1.0], saturation in the range [0.5, 1.0], and a hue delta of 5. At last, the results were normalized using precomputed per-channel dataset statistics listed in Section 3.2.

The results of our experiments are presented in the Table 4 in supplementary material. All the models trained on our dataset achieve high metrics, with the FPN model outperforming others in both tasks.

5.2 Cross-Dataset Evaluation

We conduct cross-dataset evaluation to compare our dataset with existing ones in face parsing and portrait segmentation domains.

Experiments Configuration. We train the FPN model for two segmentation tasks on each dataset. All datasets' samples were exposed to resizing to fixed 384 × 384 shape and base augmentations pipeline described in Section 5.1. The training process and model configuration are the same as the base experiments.

Table 2: Cross-dataset evaluation results. Each cell value contains mean IoU (mIoU) metrics for the corresponding training and testing sets pair. Train (test) average mIoU represents the overall mIoU value on the listed testing (training) sets. The high train average mIoU metric directly relates to the dataset's generalization ability. A low test average mIoU metric reflects dataset complexity, as a model pre-trained on a different set struggles to achieve a high metric. We highlighted the best metric in each column to emphasize the dataset's ability to generalize to other distributions. The best metric in all columns except the last one was chosen, excluding diagonal values.

Portrait Segmentation						
		Tested				
	Dataset	EasyPortrait (ours)	FVS	HumanSeg14K	Face Synthetics	Train avg. mIoU
Trained	EasyPortrait (ours)	98.64	**97.86**	**93.18**	**97.76**	**96.86**
	FVS (Kuang and Tie, 2021)	79.05	96.24	90.6	80.36	86.56
	HumanSeg14K (Chu et al., 2021)	76.01	96.23	97.53	71.66	85.35
	Face Synthetics (Wood et al., 2021)	**84.99**	57.14	57.87	99.44	74.86
	Test avg. mIoU	84.67	86.87	84.8	87.31	

Face Parsing							
		Tested					
	Dataset	EasyPortrait (ours)	CelebAMask-HQ	iBugMask	Face Synthetics	LaPa	Train avg. mIoU
Trained	EasyPortrait (ours)	81.51	76.01	39.0	**51.2**	61.03	61.75
	CelebAMask-HQ (Lee et al., 2019)	66.17	83.41	**54.74**	46.6	60.62	62.31
	iBugMask (Lin et al., 2021)	61.58	**79.1**	64.59	44.42	**66.3**	63.19
	Face Synthetics (Wood et al., 2021)	55.55	40.67	18.84	83.12	42.63	48.16
	LaPa (Liu et al., 2020)	**68.56**	73.92	47.66	48.05	79.02	**63.44**
	Test avg. mIoU	66.67	70.62	44.97	54.68	61.92	

Portrait Segmentation. Besides our dataset, the model was trained and tested on HumanSeg14K (Chu et al., 2021), Face Synthetics (Wood et al., 2021), and FVS (Kuang and Tie, 2021) portrait segmentation datasets. We could not include the EG-1800 (Shen et al., 2016) and the AiSeg (ais, 2019) datasets due to a lack of images on the public shared sources and inappropriate samples, respectively.

Some additional preprocessing steps were applied:

- We led the EasyPortrait's class "person" to a consistent appearance by labeling others classes (without "background") as "person" class.
- FVS (Kuang and Tie, 2021) contains non-binary masks, necessitating binarization. Pixel values clustered near 0 or 255, so we used a threshold of 127 to separate "person" and "background". The dataset was split into 1,326 training and 935 testing samples, with 200 images randomly selected from the training set for validation.
- HumanSeg14K (Chu et al., 2021) dataset was divided into the training, validation, and test parts with 8,770, 2,431, and 2,482 samples, respectively.
- Similar to EasyPortrait's preprocessing, we prepared the Face Synthetics (Chu et al., 2021) dataset to portrait segmentation masks. We randomly picked 75,000 training, 15,000 testing, and 10,000 validation samples.

Face Parsing. Given that the EasyPortrait skin class was annotated using unique rules and most face parsing datasets lack annotations for the teeth class, we selected only six classes for cross-dataset evaluation: "background", "left brow", "right brow", "left eye", "right eye", "lips". We adopted the original annotations of face parsing datasets to the target ones:

- Such EasyPortrait classes as "teeth", "person" and "skin" classes were mapped to the "background".
- Since lips of the CelebAMask-HQ dataset are divided into two classes: "lower lip" and "upper lip", we combined them into one "lips" class. The remaining classes are considered as background. All datasets below were preprocessed similarly to CelebAMask-HQ (Lee et al., 2019). We divided the CelebAMask-HQ dataset into 22,500 training, 3,000 validation, and 4,500 test samples.
- The LaPa (Liu et al., 2020) dataset was originally split into 18,167 training, 2,000 validation, and 2,000 test samples.
- Originally, images from iBugMask (Lin et al., 2021) were split into 21,866 training and 1,000 testing examples. The validation set was randomly sampled from the training set and contained 1,866 images. Note that iBugMask (Lin et al., 2021) contained images with a bounding box for the face in the provided mask. To avoid parsing other faces, we crop them as described in the original paper.
- The Face Synthetics dataset was distributed into 75,000 training, 10,000 validation, and 15,000 test samples by us.

Results. The cross-dataset evaluation results in Table 2 demonstrate that the EasyPortrait has the best generalization capability regarding mIoU metrics on each portrait segmentation test set. Also, the FPN model, trained on the EasyPortrait dataset for portrait segmentation, surpasses FVS results even on their own test set. Due to the reduced list of classes, the quantitative assessment provides limited insights into the dataset's applicability for the face parsing task. Besides, the videoconferencing domain is slightly limited: images in other datasets are more heterogeneous in context, displaying multiple people and different activities, while EasyPortrait consistently shows a single person in front of a computer or phone. Despite this, we achieved concurrent results in the face parsing task.

6 DISCUSSION

Ethical Considerations. Creating facial feature datasets involves significant ethical considerations. Individuals must provide informed consent for their facial data to be collected, stored, and used. Instructions for the crowd tasks were presented in clear language. Additionally, ensuring diversity in the dataset is vital to prevent biases leading to unfair treatment of certain groups. To address privacy concerns, all crowd workers signed a consent form allowing us to process and publish their photos. We adhere to the Federal Law ”On Personal Data” (27.07.2006 N152) of Russia, ensuring legal compliance in data handling. Besides, we used anonymized user ID hashes to protect crowd workers' privacy. Despite the limitations of using the Russian crowdsourcing platform, efforts were made to minimize biases and make the dataset as racially diverse as possible. Thus, the three most frequently identified races – Caucasian, Negroid, and Mongoloid – are covered.

Possible Misuse. Facial feature datasets can be misused in various ways, such as enhancing surveillance systems for mass tracking and profiling individuals based on race or other attributes, leading to discrimination. They also pose risks of creating deepfakes and identity theft. Note that we employ the collected data exclusively for research purposes and urge the potential users to follow ethical research practices, which include proper citation, acknowledgment of data sources, and sharing their work under the same license as the original.

Limitations. The proposed dataset was designed for one concrete domain — video conferencing, which entailed some limitations in scenes, occlusions, and the number of people in photos. While resistance to the multiplicity of subjects in inference can be achieved by mosaic augmentation, the system based on EasyPortrait can produce errors on unspecific backgrounds and accessories.

Another area for improvement in the current version of EasyPortrait is the lack of some classes, which reduces the number of its applications. However, we plan to significantly expand the classes by “mouth”, “hair”, “headphones”, “glasses”, “earrings”, “nose”, “hat”, “neck”, and “beard”.

The limited choice of crowdsourcing platforms caused a bias towards the Caucasian race. Despite the attempts to diversify the data regarding subjects' countries, the bias toward Russia remained (see Fig. 3a). Analogically, the balance between subjects' emotions and their heads' turns was not fully reached (see Fig. 3i and Fig. 8, respectively). Overcoming the described limitations presents an opportunity to develop a system that will be more robust to variability in the environment and people. Expanding dataset diversity will be a valuable focus for future work.

7 CONCLUSION

We propose a large-scale image dataset for portrait segmentation and face parsing, which contains 40,000 indoor photos of people, each with a high-quality 9-class semantic mask. Easyportrait supports beautification and segmentation tasks like background removal, skin enhancement, and teeth whitening, enhancing user experience in video conferencing apps. We performed extensive model experiments and cross-dataset comparisons, with an ablation study highlighting the importance of data quantity and head pose diversity for robust training. Future work includes adding occlusions to annotations and expanding the number of classes (see Section 6).

REFERENCES

(2019). AISeg. https://github.com/aisegmentcn/matting_human_datasets.

(2020). Persons Labeled. https://ecosystem.supervise.ly/projects/persons.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation.

Chu, L., Liu, Y., Wu, Z., Tang, S., Chen, G., Hao, Y., Peng, J., Yu, Z., Chen, Z., Lai, B., and Xiong, H. (2021). Pp-humanseg: Connectivity-aware portrait segmentation with a large-scale teleconferencing video dataset. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 202–209.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation.

Guo, J., Zhu, X., and Lei, Z. (2018). 3ddfa. https://github.com/cleardusk/3DDFA.

Kae, A., Sohn, K., Lee, H., and Learned-Miller, E. G. (2013). Augmenting crfs with boltzmann machine shape priors for image labeling. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026.

Kuang, Z. and Tie, X. (2021). Flow-based video segmentation for human head and shoulders. *ArXiv*, abs/2104.09752.

Le, V., Brandt, J., Lin, Z. L., Bourdev, L. D., and Huang, T. S. (2012). Interactive facial feature localization. In *European Conference on Computer Vision*.

Lee, C.-H., Liu, Z., Wu, L., and Luo, P. (2019). Maskgan: Towards diverse and interactive facial image manipulation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5548–5557.

Li, J., Ma, S., Zhang, J., and Tao, D. (2021). Privacy-preserving portrait matting. *Proceedings of the 29th ACM International Conference on Multimedia*.

Liang, J., Zeng, H., Cui, M., Xie, X., and Zhang, L. (2021). Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 653–661.

Lin, S., Ryabtsev, A., Sengupta, S., Curless, B., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2020). Real-time high-resolution background matting. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8758–8767.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection.

Lin, Y., Shen, J., Wang, Y., and Pantic, M. (2021). Roi tanh-polar transformer network for face parsing in the wild. *Image Vis. Comput.*, 112:104190.

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 *AAAI Conference on Artificial Intelligence*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild.

Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation. *2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431-3440.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *2019 The International Conference on Learning Representations (ICLR)*, abs/1711.05101.

Luo, L., Xue, D., and Feng, X. (2020). Ehanet: An effective hierarchical aggregation network for face parsing. *Applied Sciences*, 10(9):3135.

Park, H., Sjösund, L. L., Monet, N., Yoo, Y., and Kwak, N. (2019a). Sinet: Extreme lightweight portrait segmentation networks with spatial squeeze modules and information blocking decoder. *arXiv preprint arXiv:1911.09099*.

Park, H., Sjösund, L. L., Yoo, Y., and Kwak, N. (2019b). Extremec3net: Extreme lightweight portrait segmentation networks using advanced c3-modules. *arXiv preprint arXiv:1908.03093*.

Poudel, R. P. K., Liwicki, S., and Cipolla, R. (2019). Fast-scnn: Fast semantic segmentation network.

Ryumina, E., Dresvyanskiy, D., and Karpov, A. (2022). In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*.

Sander, E. L. J. (2020). Coronavirus could spark a revolution in working from home: Are we ready? *The conversation*.

Shen, X., Hertzmann, A., Jia, J., Paris, S., Price, B. L., Shechtman, E., and Sachs, I. (2016). Automatic portrait segmentation for image stylization. *Computer Graphics Forum*, 35.

Taherkhani, F., Nasrabadi, N. M., and Dawson, J. (2018). A deep face identification network enhanced by facial attributes prediction.

Wood, E., Baltruvsaitis, T., Hewitt, C., Dziadzio, S., Johnson, M., Estellers, V., Cashman, T. J., and Shotton, J. (2021). Fake it till you make it: face analysis in the wild using synthetic data alone. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3661–3671.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems*.

Yin, X. and Chen, L. L. (2022). Faceocc: A diverse, high-quality face occlusion dataset for human face extraction. *ArXiv*, abs/2201.08425.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *European Conference on Computer Vision*.

Zhu, X., Liu, X., Lei, Z., and Li, S. Z. (2017). Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*.

APPENDIX

In the supplementary materials, we include illustrations of selected samples with their corresponding annotation masks from the EasyPortrait dataset, the head turns distributions across different datasets, the guidelines for class annotations, and the evaluation results on the EasyPortrait dataset.

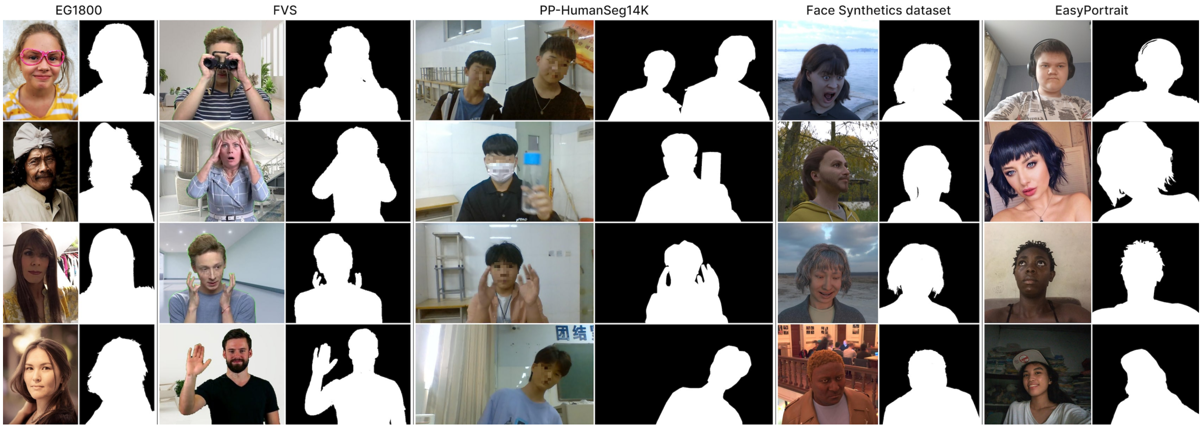

Figure 5: Visual comparison of existing portrait segmentation datasets. One can notice high-frequency details (e.g., hair) in segmentation masks in samples from our dataset. The AiSeg (ais, 2019) dataset is not considered since it provides the extracted foreground images without a corresponding annotation mask.

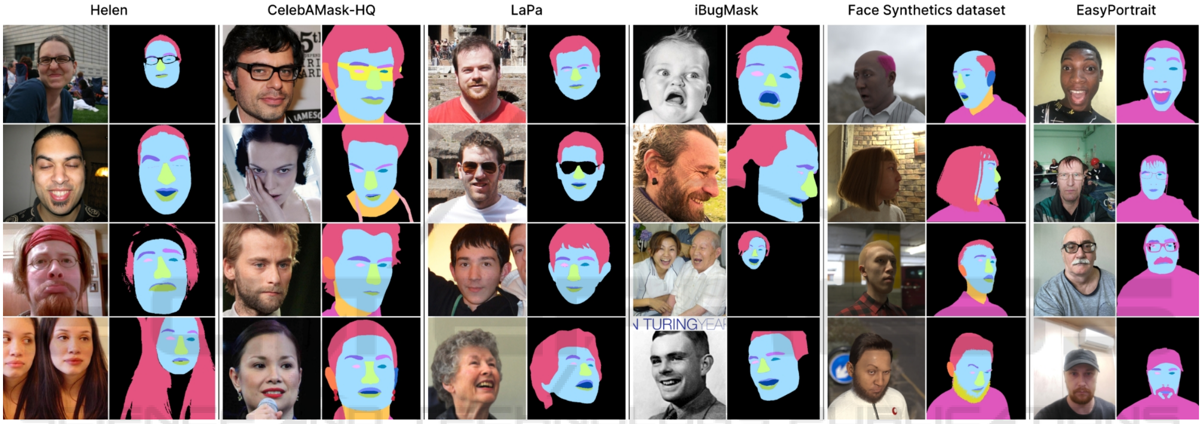

Figure 6: Visual comparison of existing face parsing datasets. Only Face Synthetics (Wood et al., 2021) and EasyPortrait datasets can be used to solve background removal and face enhancement problems. None of them except EasyPortrait can be used for teeth whitening. We do not include LFW-PL (Kae et al., 2013) and FaceOcc (Yin and Chen, 2022) datasets in the visualization due to the lack of classes and the need for preprocessing, respectively.

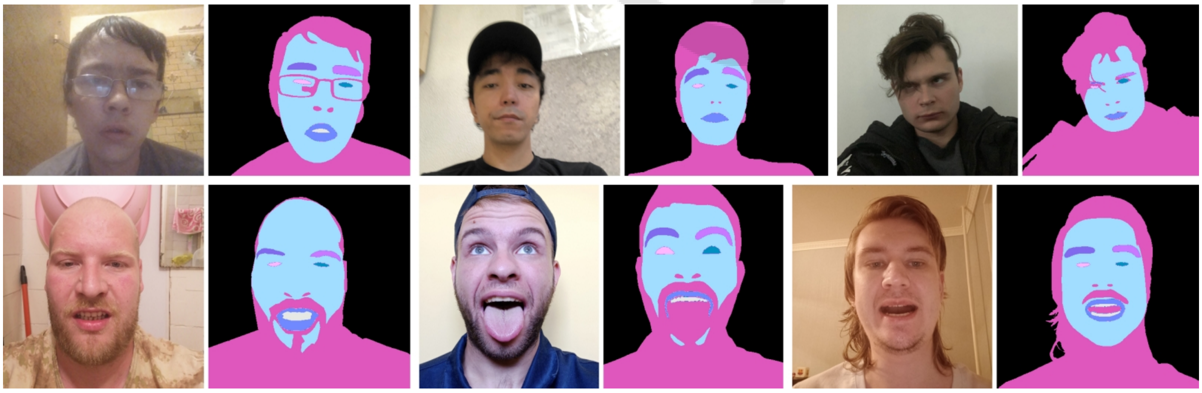

Figure 7: Visualization of the beard annotation rules. (up) The beard is included in the skin if it is a separate hair or barely noticeable. (bottom) The beard is excluded from the skin if it is clear.

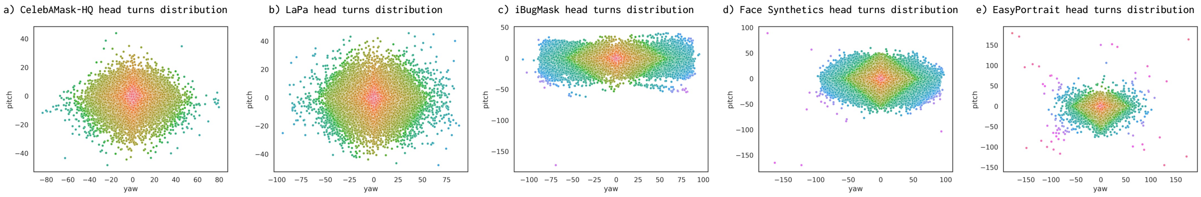

Figure 8: Head turns distributions for several face parsing and portrait segmentation datasets, including EasyPortrait. Yaw and pitch coefficients were obtained via 3DDFA network (Guo et al., 2018).

Table 3: EasyPortrait annotators rules.

Class	Rules
Person	– headphones and things in front of the person are defined as a person's class – individual hairs and all empty areas closed by a person are not included in the person class
Eyebrows	stand out along a strict border, excluding individual hairs
Eyes	distinguished by whites, excluding eyelids and eyelashes
Skin	– the skin class should affect only skin without hair, eyes, and other face attributes – the boundaries of the skin of the face or person should be highlighted logically on overexposed or darkened photos – the rare bristle also considered skin – ears, second chin, and nostrils are not included in the skin class
Teeth	teeth and everything else in the open mouth stand out separately, the latter as an occlusion
Occlusions	– makeups and piercing are defined as occlusions – the part of eyeglasses which cover skin should be annotated as occlusion, including sunglasses and glare on clear glasses – beard with a strict border are considered occlusion – the tongue out of the mouth should be annotated as occlusion

Table 4: Evaluation results on the EasyPortrait. Column "mIoU" is divided into two subcolumns: face parsing and portrait segmentation tasks. For face parsing, we present mIoU metrics for each class separately, while for portrait segmentation, we provide only the overall mIoU score. We additionally trained FPN and Segformer-B0 on 224 × 224, 512 × 512, and 1024 × 1024 resolutions to demonstrate the overall increasing tendency amongst both convolutional and transformer models depending on increasing resolution.

Model	Input Size	Model Size (MB)	FPS	mIoU								
				Face Parsing								PS
				skin	l-eye	r-eye	l-brow	r-brow	lips	teeth	overall	overall
BiSeNetv2 (Yu et al., 2018)	384	56.5	91.47	90.75	71.94	72.57	67.67	67.53	80.87	63.09	76.72	97.95
SegFormer-B0 (Xie et al., 2021)		14.19	72.45	92.05	78.55	79.26	72.5	72.21	83.53	73.52	81.38	98.61
FCN + MobileNetv2 (Long et al., 2014)		31.17	66.07	90.49	69.95	70.63	66.29	66.09	79.23	59.84	75.23	98.19
FPN + ResNet50 (Lin et al., 2017)		108.91	58.1	**92.28**	**79.48**	**80.08**	**72.64**	**72.47**	**84.15**	**74.09**	**81.83**	**98.64**
DeepLabv3 (Chen et al., 2017)		260.02	25.65	91.77	73.78	74.63	69.61	69.74	83.42	70.53	79.11	98.63
Fast SCNN (Poudel et al., 2019)		6.13	93.89	88.58	58.42	58.7	58.68	58.87	73.16	44.86	67.56	97.64
DANet (Fu et al., 2019)		190.29	42.43	91.8	74.01	74.93	70.01	69.75	83.7	70.8	79.3	98.63
EHANet (Luo et al., 2020)		44.81	132.78	89.68	68.87	69.26	63.6	63.82	73.98	52.05	72.56	-
SINet (Park et al., 2019a)		0.13	134.18	-	-	-	-	-	-	-	-	93.32
ExtremeC3Net (Park et al., 2019b)		0.15	71.75	-	-	-	-	-	-	-	-	96.54
SegFormer-B0	224	14.9	74.84	90.19	68.59	70.46	65.79	65.72	77.94	60.66	74.83	98.17
FPN + ResNet50		108.91	61.56	90.6	69.67	71.88	65.84	65.64	78.94	62.95	**75.6**	**98.31**
SegFormer-B0	512	14.9	65.88	92.5	81.03	81.18	74.31	74.08	84.87	78.14	83.19	**98.66**
FPN + ResNet50		108.91	53.14	92.55	81.55	81.47	74.33	74.38	85.27	77.77	**83.33**	98.64
SegFormer-B0	1024	14.9	62.9	93.13	84.2	83.97	76.41	76.12	86.88	83.2	**85.42**	**98.74**
FPN + ResNet50		108.91	52.34	92.94	84.55	84.24	76.11	76.11	86.93	82.62	85.37	98.54