Journal of Ambient Intelligence and Humanized Computing (2023) 14:9383–9402
https://doi.org/10.1007/s12652-023-04607-8

ORIGINAL RESEARCH

Deep face segmentation for improved heart and respiratory rate estimation from videos

Marc-André Fiedler[1] · **Philipp Werner**[1] · **Michał Rapczyński**[1] · **Ayoub Al-Hamadi**[1]

Received: 22 June 2022 / Accepted: 4 April 2023 / Published online: 23 May 2023

Abstract
The selection of a suitable region of interest (ROI) is of great importance in camera-based vital signs estimation, as it represents the first step in the processing pipeline. Since all further processing relies on the quality of the signal extracted from the ROI, the tracking of this area is decisive for the performance of the overall algorithm. To overcome the limitations of classical approaches for the ROI, such as partial occlusions or illumination variations, a custom neural network for pixel-precise face segmentation called FaSeNet was developed. It achieves better segmentation results on two datasets compared to state-of-the-art architectures while maintaining high execution efficiency. Furthermore, the Matthews Correlation Coefficient was proposed as a loss function providing a better fitting of the network weights than commonly applied losses in the field of multi-class segmentation. In an extensive evaluation with a variety of algorithms for vital signs estimation, our FaSeNet was able to achieve better results in both heart and respiratory rate estimation. Thus, a ROI for vital signs estimation could be created that is superior to other approaches.

Keywords Camera-based monitoring · Heart rate · Remote photoplethysmography · Respiratory rate · Vital signs

1 Introduction

In recent years, considerable efforts have been invested by the research community to further enhance methods for camera-based estimation of vital signs. Since it was discovered in 2008 that subtle color changes can be measured in human skin pixels of frames from video sequences based on the principle of photoplethysmography (PPG) (Verkruysse et al. 2008), great progress has already been made in the development of these techniques. These color changes occur due to cardiac-synchronous variations in the amount of reflected light, which are caused by the changing blood volume in the arteries with every heartbeat (Poh et al. 2010). In turn, these cardiac-synchronous variations are linked to respiration via the phenomenon of respiratory sinus arrhythmia, which results in a rise of heart frequency during inhalation and a fall during exhalation (Zhao et al. 1994). Thus, remote PPG (rPPG) systems can be used to capture important health-related information to assess and diagnose cardiovascular diseases (Castaneda et al. 2018). From the obtained rPPG signal, two of the most important vital signs, namely the heart rate (HR) and the respiratory rate (RR), can be calculated (Elliott and Coventry 2012). Both are important biomarkers for the prevention and diagnostics of various illnesses (Moraes et al. 2018). For example, the HR is a measure of physiological activity and has the ability to indicate a person's state of health (Fel and Malik 1994). The RR, for instance, can be used to determine whether a patient is in critical condition and is therefore part of many risk scores (Becker et al. 2017).

The processing pipeline of algorithms for vital signs estimation can usually be divided into three main processing steps: Firstly, a suitable Region of Interest (ROI) has to be selected and tracked across each frame of a video sequence. Subsequently, a rPPG signal can be generated by averaging the color values from the ROIs, possibly with the help of some signal processing methods. Finally, the vital signs can be estimated from the obtained signal. A general schematic diagram of this processing pipeline is shown in Fig. 1. As the detection of the ROI is the first of these stages, it is crucial for the later performance of the overall system. In this

✉ Marc-André Fiedler
marc-andre.fiedler@ovgu.de

[1] Neuro-Information Technology Group, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany

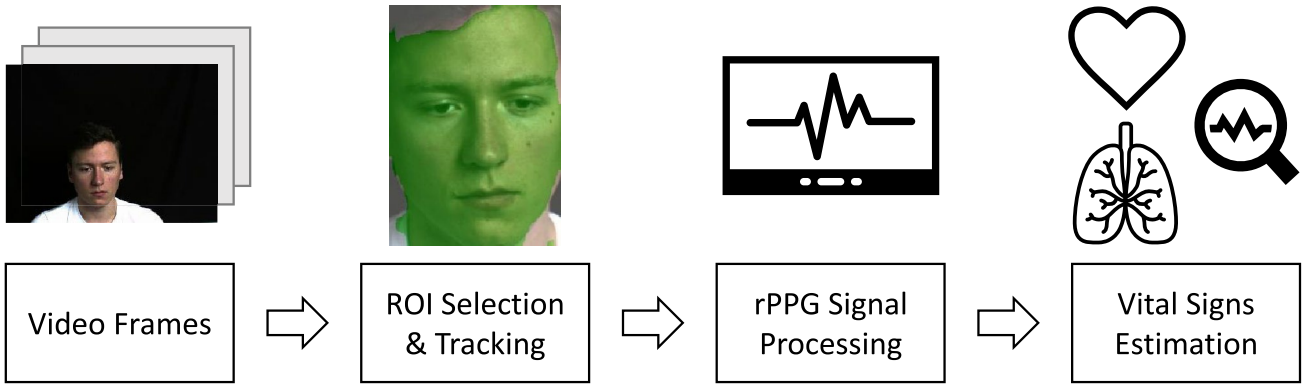


Fig. 1 General schematic pipeline of algorithms for vital signs estimation in remote photoplethysmography (rPPG)

work, we focus on approaches that are using as ROI (partial) areas of the face, since we believe that they have the greatest potential for future applications. This is for two reasons: First, the face exhibits significantly higher PPG signal energies than other body regions due to the high density of blood vessels and the relatively thin layer of skin (Nilsson et al. 2007). Second, the region needs to be permanently visible and uncovered by clothing, making the face the most suitable body area for this purpose.

The goal in selecting the ROI is to capture as many skin pixels as possible with only a very small percentage of non-skin pixels, because they will create undesirable artifacts in the rPPG signal. If the ROI is chosen too small, the quantization noise of the cameras may not be sufficiently attenuated by averaging the pixel intensities, which is why larger ROIs often have positive impact on the signal-to-noise ratio (Wang et al. 2018). In order to define the ROI accurately, most rPPG algorithms employ a face detection module upstream and subsequently localize the ROI on the basis of the resulting face crop. Most methods in prior work use static geometric regions as ROI based on the face bounding box and/or facial landmarks, such as the forehead (Sanyal and Nundy 2018; Blöcher et al. 2017) or cheeks (Feng et al. 2015; Nisar et al. 2016), or utilize pixel-based skin segmentation with color thresholds inside the face crop (Rapczynski et al. 2018; Wang et al. 2017b; Fouad et al. 2019). However, predefined static ROIs have the disadvantage that they are unable to react to interfering pixels caused by hair, beard, glasses, headgear and others. This can have significant negative effects on the estimation performance. Threshold-based skin segmentation methods can deal with this type of problem through pixelwise skin/non-skin mapping, but will fail under skin tone variations (e.g. for strong redness) and changing, too weak or too strong ambient illumination scenarios.

To overcome those limitations of previous ROI approaches, we present in this paper a framework for deep face segmentation to improve the quality of the rPPG signal generated out of video frames and to enable better estimations for the vital signs HR and RR. By segmenting all pixels of the image into the three classes face, hair and background and by taking advantage of semantic information, we aim to achieve pixel-precise segmentation results even in case of partial face occlusions and illumination variations. This should guarantee that only pixels containing photoplethysmographic information are included in the computation of the rPPG signal. In addition, we take care to ensure that the convolutional neural network (CNN) employed maintains high execution efficiency, allowing face mask generation to be performed in real-time on a conventional graphics processing unit (GPU) and thus to be suitable for usage in real-world applications. The contributions of this paper can be summarized as follows:

1. We propose a custom CNN architecture for real-time face segmentation that outperforms other state-of-the-art models in terms of segmentation performance while maintaining high execution efficiency.
2. A novel approach for the loss function in multi-class semantic segmentation was developed. It is based on the Matthews Correlation Coefficient, which is particularly well suited for handling imbalanced datasets.
3. A modular framework for rPPG algorithms was created, where the individual processing steps of the pipeline (see Fig. 1) can be easily interchanged.
4. Comprehensive experimental validation was performed by comparing different CNN architectures for semantic segmentation in terms of performance and execution efficiency, as well as evaluating the impact of the application for multiple ROIs in several vital signs estimation algorithms for HR and RR.

Our paper is structured in the following way: In Sect. 2, methods are described outlining our newly proposed CNN architecture, the designed loss function, the applied training details, and the generation of the rPPG signal. Subsequently, the experiments are reported in Sect. 3 specifying the used

datasets, evaluation metrics, and state-of-the-art benchmarking algorithms. In Sect. 4, the experimental results are presented and discussed. Finally, a conclusion is drawn in Sect. 5.

2 Methods

In this section, our newly proposed face segmentation network (**FaSeNet**) is presented in detail. It is designed to accurately segment faces of humans in video frames to facilitate most precise estimations of their vital signs. First, the network architecture is introduced. Next, we elaborate the novel loss function employed and the details of our training. The generation of rPPG signals from the segmented face masks is outlined at the end of this section.

2.1 Network architecture

The architecture of our FaSeNet consists of the fusion of two encoder branches for feature extraction and the remaining decoder layers for spatial recovery of the input image resolution. The two paths for encoding are referred to as spatial branch and context branch. The goal of running the two branches in parallel is to produce high quality feature maps with shallow spatial detail as well as global contextual information, without having a huge negative impact on the inference speed of the network. Our overall architecture is shown in Fig. 2.

The aim of the spatial branch is to encode rich spatial information. We try to achieve this by not downsampling the feature maps excessively compared to the original resolution of the input image. For this purpose, a cascade of four custom-designed blocks is adopted. Each of them is made up of a depthwise convolution and a later convolutional layer, both followed by batch normalization and rectified linear unit (ReLU) activation. This design was inspired by MobileNet (Howard et al. 2017). Batch normalization is adopted after convolution and before activation to improve generalization while accelerating training. ReLU is then employed to enhance the nonlinear fitting ability of the model. Both types of convolutions use kernel size 3×3. We refer to this block as "DepthConv + Conv" and plot it in Fig. 3. The advantage of depthwise convolution is that it reduces the number of parameters, thus lowering the computational cost while maintaining similar performance. Therefore, a higher number of filters and a stride = 1 are used for this depthwise convolution in order to increase modeling capacity. Compared to the later convolution the filter amount is tripled (48, 96, 192 and 384). The downsampling is performed by the subsequent convolutional layers with less filters (16, 32, 64 and 128) and stride = 2. By combining depthwise convolutions with many filters and convolutions for downsampling with few

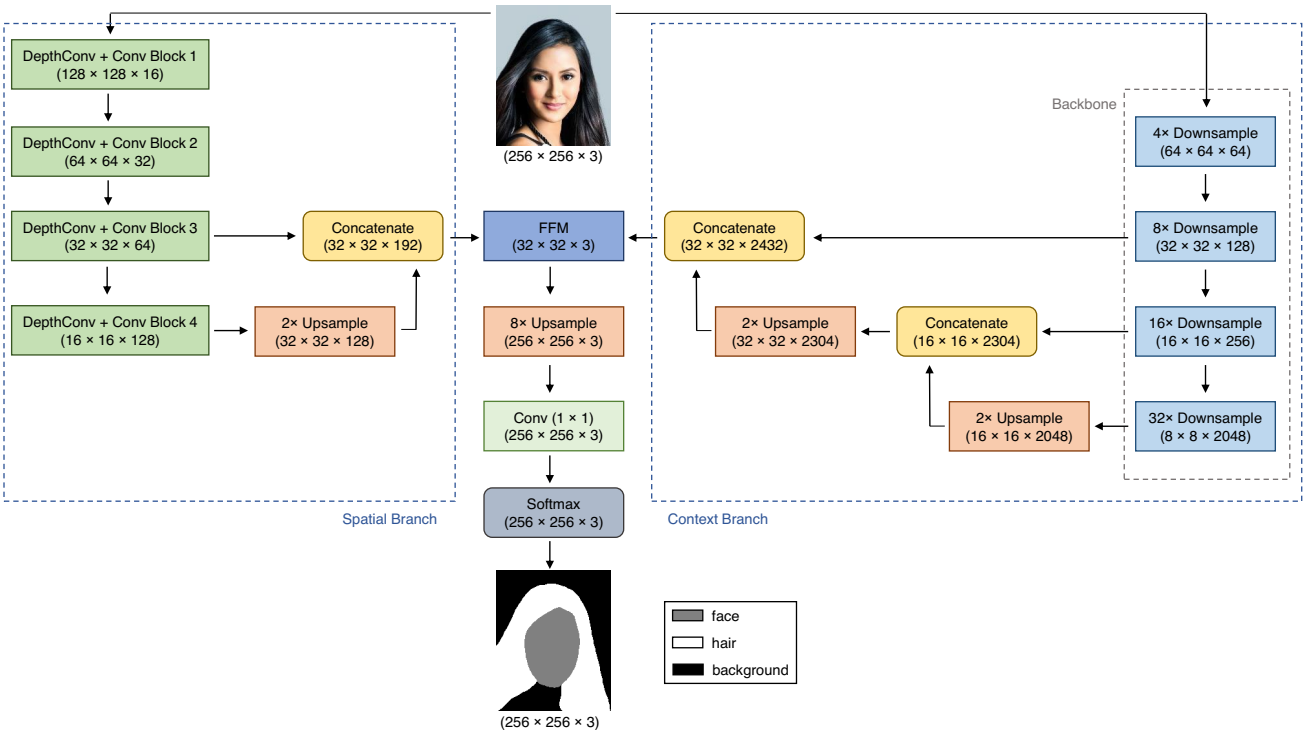


Fig. 2 Network architecture of FaSeNet: Each block describes one or more layers and specifies the corresponding output feature map size

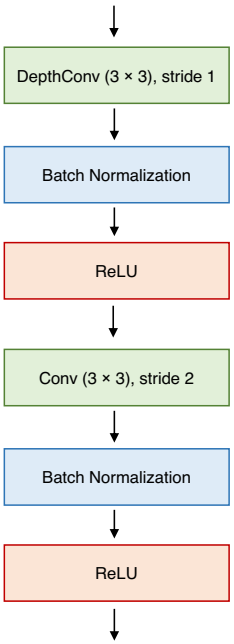


Fig. 3 "DepthConv + Conv" Block of FaSeNet

filters, better run-time properties can be generated for the network architecture without reducing the quality of the features. The output of the last of these four "DepthConv + Conv" blocks is upsampled and concatenated with the one of the third. This branch structure yields high quality spatial information at manageable computational cost.

The context branch is employed for providing a large receptive field to encode valuable semantic context information. For this purpose, we apply ResNet50V2 (He et al. 2016b) as lightweight backbone for downsampling the feature maps. The weights are initialized with values trained on ImageNet (Deng et al. 2009). ResNet50V2 was chosen to be the backbone because of its excellent run-time characteristics in combination with its high performance capability. A schematic representation of the ResNet50V2 backbone with labels for the individual blocks and layers can be found in Fig. 4. At the tail of the backbone with the maximum receptive field, the resulting feature map is upsampled and concatenated with the output of layer "conv4_block6_1_relu". For ResNets (He et al. 2016a), each convolutional layer is followed by a batch normalization and a ReLU subsequently. The feature maps are tapped after non-linear activation. This procedure is repeated accordingly to concatenate the feature map with output of layer "conv3_block4_1_relu". Thus, the high-level feature maps from different dimensions are combined and upsampled accordingly to enhance the model's capability.

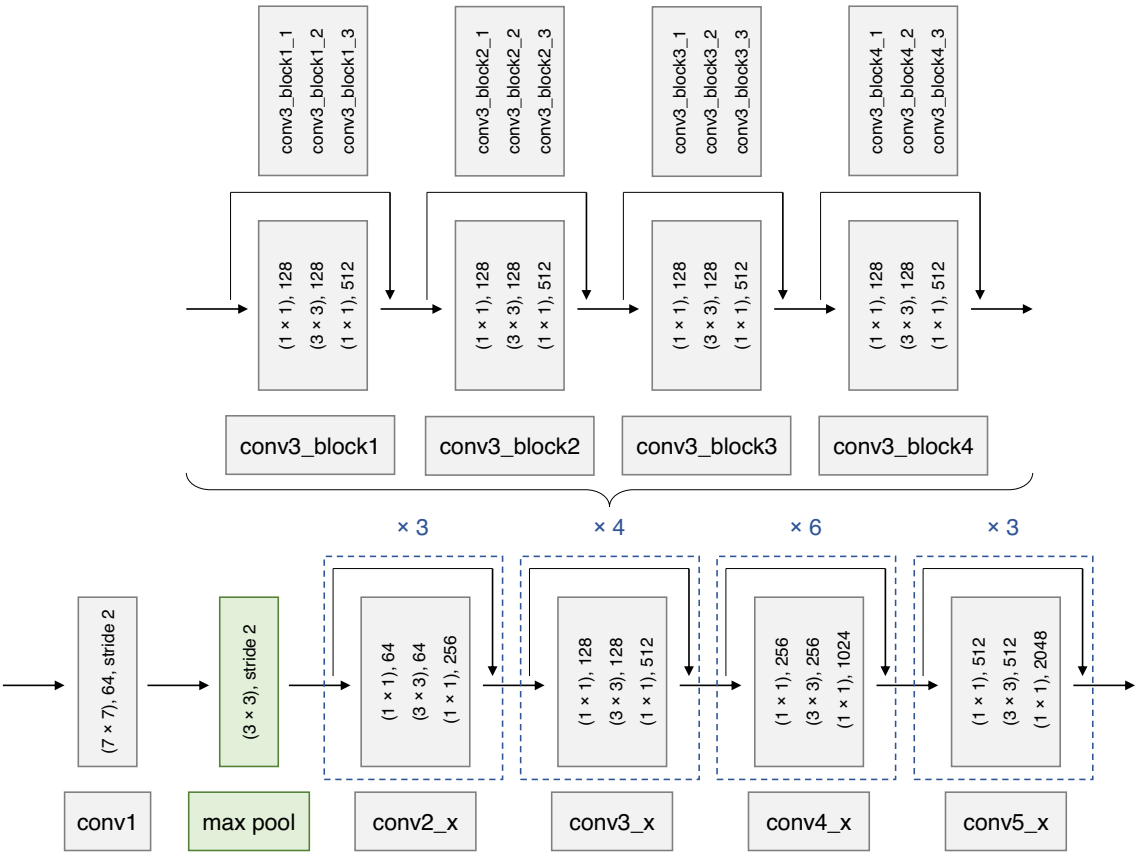


Fig. 4 Schematic representation of the ResNet50V2 backbone with labels for the individual blocks and layers

The combination of good spatial information and sufficient receptive field are crucial for achieving high performance in semantic segmentation. To accomplish this within our model by merging the two encoder branches, we employ the feature fusion module (FFM) created by Yu et al. (2018). It was explicitly designed to fuse different levels of feature representation, while focusing on important features and suppressing unimportant ones. For this purpose, the two input branches are first concatenated and processed by a layer consisting of convolution, batch normalization and ReLU. This convolution combines the information and condenses it to three channels. Subsequently, a global pooling layer and two convolutional layers with ReLU and sigmoid activation, respectively, are run through for computing an attention weight vector. In addition, skip connections are used, which allow information flow between feature hierarchies (Ulku and Akagündüz 2022). These additional paths are beneficial for model convergence. Thus, the FFM input feature maps are linked to the outputs via multiplication and addition, respectively.

Following the FFM module, the decoder branch begins, which is responsible for restoring the original image resolution from the beginning. Only details for fine-tuning are learned during this process, the intrinsic features originate from the encoder. For this reason, fast bilinear upsampling with a factor of 8 takes place to save computational expenses. Then, pointwise 1×1 convolution is performed for feature pooling. At the end, softmax is applied for generating the output class probabilities.

2.2 Loss function

The particular choice of the loss function has a great impact on the later performance of the Deep Learning model, as it needs to ensure that the objective is learned accurately and rapidly. In semantic segmentation, the main distinction is between distribution-based and region-based loss functions (Jadon 2020). In particular, when choosing a loss for pixel-based classification, the problem of imbalanced class distribution must be considered. The analysis of literature surveys shows that (distribution-based) Cross-Entropy and (region-based) Dice loss functions are mostly used for training models in semantic segmentation (Garcia-Garcia et al. 2018).

In this paper, we propose a novel approach for the loss function by calculating it based on the Matthews Correlation Coefficient (**MCC**). To the best of our knowledge, we are the first to introduce MCC loss ($\mathbf{L_{MCC}}$) in multi-class semantic segmentation and classification. It has been shown in other works that it is a more reliable metric for evaluating classification performance (Chicco and Jurman 2020; Chicco et al. 2021) and particularly suitable for handling data imbalance (Boughorbel et al. 2017; Zhu 2020). It is generally considered as a balanced measure that can be used even with classes of very different sizes (Zhu 2020), and has been mainly employed in various bioinformatics applications as a performance metric for classifiers in order to guarantee better generalization results (Song et al. 2006; Wang et al. 2015; Huang et al. 2010). Compared to Dice loss, the advantage of MCC is that it involves the number of true negative samples in the calculation. Additionally, the eight most relevant derived ratios obtained by combining all components of a confusion matrix are integrated together in MCC (Lever et al. 2016).

MCC is defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{\begin{matrix}(TP + FP)(TP + FN) \\ (TN + FP)(TN + FN)\end{matrix}}} \tag{1}$$

indicating the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predicted pixels.

Since we apply this loss to a multi-class problem, we compute the mean MCC (**mMCC**) as follows:

$$mMCC = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{\begin{matrix}(TP_i + FP_i)(TP_i + FN_i) \\ (TN_i + FP_i)(TN_i + FN_i)\end{matrix}}} \tag{2}$$

where

$$TP_i = p_{ii} \tag{3}$$

$$TN_i = \sum_{j=1}^{N} p_{jj}, \quad j \neq i \tag{4}$$

$$FP_i = \sum_{j=1}^{N} p_{ji}, \quad j \neq i \tag{5}$$

$$FN_i = \sum_{j=1}^{N} p_{ij}, \quad j \neq i \tag{6}$$

Let p_{ij} be the total number of pixels belonging to class i but classified to class j and N the total number of classes.

The final loss function is thus obtained as follows:

$$L_{MCC}(y, \hat{y}) = 1 - mMCC(y, \hat{y}) \tag{7}$$

where $\hat{y}$ denotes the model prediction and y the ground truth.

2.3 Training details

For training our FaSeNet model, images and masks are zero padded to be square and resized to 256×256 pixels,

if this is not already the case. As training data is limited, several data augmentation operations are conducted on the inputs: random horizontal flipping, vertical flipping with probability of 10%, random rotation in range [-30°, 30°], random horizontal and vertical shifting in range [-20%, 20%] of the image, and adjusting brightness, contrast, hue and saturation. Details on the training datasets can be found in Sect. 3.1.1.

All training is performed on a NVIDIA RTX 2080 Ti using TensorFlow framework with Python. The batch size is set to 32. Our model is trained for 300 epochs in total. As optimizer stochastic gradient descent (SGD) is employed with momentum of 0.98. The "poly" learning rate policy is utilized by multiplying the initial learning rate by $(1 - \frac{iter}{max\_iter})^{power}$. The initial learning rate is set to 0.0025 and the power is set to 0.9. All layers for which it is possible use L2 regularization with factor 0.00002. The total training time for our final model (see Sect. 2.4) was 8 h and 23 min.

2.4 rPPG signal generation

The generation of the rPPG signal from video frames is carried out as it is standard practice in the field of rPPG. The exact procedure of the algorithm is illustrated below.

First, a face crop is determined for each frame of the video sequence by applying CNN based face detection. RetinaFace (Deng et al. 2020) is employed for this purpose. In case multiple faces are detected in a single frame, the one with the highest confidence score is selected. Subsequently, the ROI is determined within this face crop. For our FaSeNet, the crop is initially padded with zeros to become square and then resized to 256 × 256 pixels, if this is not already the case. This serves as input to our FaSeNet model, which was previously trained on the LFW-PL trainval set and the full CelebHair dataset. Details on the training process can be found in Sect. 2.3. In Sect. 3.1.1, information on the training datasets is provided. As output we obtain the predicted mask which assigns each pixel to one of the three classes. This mask is then resized to the shape of the initial face crop. Based on the resized mask, the average skin color is calculated with each frame over the entire video length for each RGB channel. Thereby, only the crop color values of face pixels are included in the calculation, the hair and background pixels are discarded. It is very important to note that the color values are taken from the initial face crop and that the face crop is not resized to the output mask shape, as resizing images can lead to the loss of important small color variations which are essential in rPPG (Rapczynski et al. 2019). For this reason, the size of the output pixel mask is adjusted and not vice versa. The resulting RGB signals are used as starting point for the HR (see Sect. 3.3.3) and RR estimation algorithms (see Sect. 3.3.4).

3 Experiments

This section describes the used datasets, evaluation metrics and benchmark algorithms in detail. The datasets and metrics are distinguished between those for face segmentation and those for camera-based vital signs estimation. In order to perform a detailed benchmark evaluation, various technical submodules and entire algorithms from the current state of the art were re-implemented to allow robust analysis and interpretation of the individual steps within the processing pipeline. In this context, lightweight CNN architectures for real-time segmentation, ROIs for generating rPPG signals from video images and algorithms for estimating HR and RR are addressed. Details about the training process can be found in Sect. 2.3.

3.1 Datasets

The datasets used for face segmentation and vital signs estimation are presented in the following subsections.

3.1.1 Datasets for face segmentation

For face segmentation two publically available datasets are used which contain segmentation masks classifying each pixel of an image into the three classes face, hair or background. Both were originally designed for hair detection and hair color recognition. However, the data is also suitable for ROI selection in rPPG systems. Especially the hair class is helpful to exclude interfering pixels caused by beard or hair hanging into the face from the generation of rPPG signals.

The Labeled Faces in the Wild Part Labels (**LFW-PL**) (Kae et al. 2013) database is an extension of the original Labeled Faces in the Wild (LFW) dataset, which provides segmentation masks for a subset of 2927 images from LFW. All images with a resolution of 250 × 250 pixels were collected from the internet in unconstrained conditions. LFW-PL data is divided into 1500 images for training, 500 for validation and 927 for testing.

The **CelebHair** (Borza et al. 2018) dataset supplies segmentation annotations for 3556 images from Large-scale CelebFaces Attributes (CelebA). The 218 × 178 images taken from celebrities in unconstrained conditions are characterized by their large head pose variations. No specifications were made by the authors for a split into training, validation and test data.

Both datasets were specifically post-processed for our application. A face detection algorithm was applied to the images and their face bounding boxes were cropped. Segmentation masks were created based on the generated bounding box locations. We only use cropped images and masks in our work for training and evaluation of the CNNs. The pixel class distribution of the cropped ground truth masks from both datasets can be found in Fig. 5.

3.1.2 Datasets for vital signs estimation

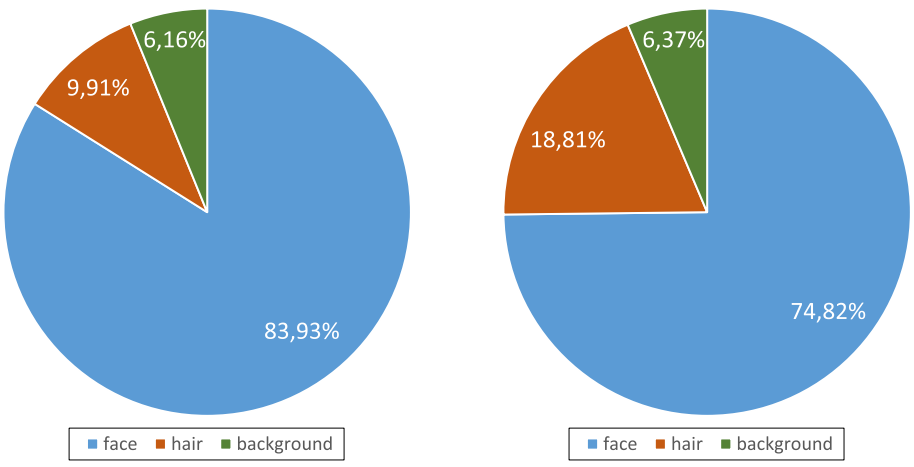
For comparison of estimation results for HR and RR, two databases for each vital sign were analyzed. Special care was taken to select only losslessly compressed video recordings, since compression removes important small changes in color information not visible to the human eye, which, nonetheless, are an elementary prerequisite for high quality vital sign estimates (Rapczynski et al. 2019).

The **MMSE-HR** dataset is a subset of the BP4D+ database published by Zhang et al. (2016) and was specifically designed to evaluate camera-based HR estimation algorithms. The subset consists of 102 video recordings of 40 subjects (23 females, 17 males), which were recorded with a resolution of 1392 × 1040 pixels at 25 frames per second (fps). The mean video length is 42.3 s. The interval of the reference HRs is [50, 128] bpm. Since the subjects performed different tasks, the video images contain large amounts of head movements.

The **PURE** (Stricker et al. 2014) data-set for HR estimation comprises 60 recordings of ten subjects (two females, eight males) while they performed the following six tasks: Steady, talking, slow translation, fast translation, small rotation and medium rotation. The mean duration of the recordings is 69.6 s. The images were recorded at 640 × 480 pixels with 30 fps and the reference HR signal was captured with a finger pulse oximeter at 60 Hz. The range of HRs lies between [44, 140] bpm. The experimental setup was illuminated by daylight through a large window, causing the ambient lighting to change over the duration of the videos due to varying cloud coverage.

For evaluation of RR algorithms, **BP4D+** (Zhang et al. 2016) is employed as one of the two databases. The dataset was originally designed for emotion recognition and thus contains many more modalities in addition to 2D videos with 1392 × 1040 pixels and respiratory signals. A total of 140 subjects (82 females, 58 males) of various ethnic ancestries were filmed during ten different tasks, resulting in 1400 videos. Since the RR signals acquired via chest belt were heavily corrupted by movements during the different tasks, a large amount of them had to be sorted out, as no ground truth could be derived with certainty due to the artifacts. At the end of this process, 269 videos with clean ground truths from 124 subjects (72 females, 52 males) remained. The video length is 68.3 s on average. The interval for the RRs is [11, 29] brpm. More details on this sorting can be found in Fiedler et al. (2020).

Additionally, we used our **own database** for RR estimation. It was first introduced in Fiedler et al. (2020) and includes videos from twelve subjects (two females, ten males). For each subject, four recordings with varying respiratory patterns were performed: Spontaneous respiration, fixed 10 breaths per minute (brpm), fixed 15 brpm and fixed 20 brpm. This results in a total of 48 videos with a resolution of 1388 × 1038 pixels at 25 fps. Each video has a uniform duration of 3 min. The respiratory pattern to perform was displayed on a monitor in front of the subjects. The ground truth was captured via a chest belt at 512 Hz. The reference RR range lies between [7, 21] brpm.



(a) LFW-PL (b) CelebHair

Fig. 5 Pixel class distribution of the cropped ground truth masks from the (**a**) LFW-PL and (**b**) CelebHair dataset

3.2 Evaluation metrics

Different metrics were employed in order to compare and validate the results obtained within the face segmentation and within the vital signs estimation. These error measures will be presented in the following for both tasks.

3.2.1 Metrics for face segmentation

The two main metrics applied for the evaluation of face segmentation are the F1 Score and the Intersection over Union (IoU). Both can also be deployed as loss function.

The **F1** Score, also known as **Dice** coefficient, is the ratio of two times the overlapping area between segmented area and ground truth of a particular class to the total value of segmented area and ground truth. For multi-class classification we calculate the mean F1 (**mF1**) as follows:

$$mF1 = \frac{2}{N} \sum_{i=1}^{N} \frac{p_{ii}}{\sum_{j=1}^{N} p_{ij} + \sum_{j=1}^{N} p_{ji}} \tag{8}$$

where p_{ij} is the total number of pixels belonging to class i but classified to class j and N is the total number of classes.

The **IoU**, often known as **Jaccard** index, is the ratio of the intersection area between predicted segmentation and ground truth to the union of these areas for every class. The mean IoU (**mIoU**) is obtained as follows:

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{p_{ii}}{\sum_{j=1}^{N} p_{ij} + \sum_{j=1}^{N} p_{ji} - p_{ii}} \tag{9}$$

In addition, the frequency weighted IoU (**fwIoU**) is given, which includes the weighting of the pixels of each class:

$$fwIoU = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}} \sum_{i=1}^{N} \frac{p_{ii} \sum_{j=1}^{N} p_{ij}}{\sum_{j=1}^{N} p_{ij} + \sum_{j=1}^{N} p_{ji} - p_{ii}} \tag{10}$$

Furthermore, the Pixel Accuracy (**PA**) which indicates the percentage of correctly classified pixels in the whole image is computed

$$PA = \frac{\sum_{i=1}^{N} p_{ii}}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}} \tag{11}$$

and the mean PA (**mPA**) value over all classes

$$mPA = \frac{1}{N} \sum_{i=1}^{N} \frac{p_{ii}}{\sum_{j=1}^{N} p_{ij}} \tag{12}$$

Since we are particularly interested in the segmentation performance for the class face, as it provides the pixels for the subsequent generation of the rPPG signal, metrics $\mathbf{F1}_{\text{face}}$, $\mathbf{IoU}_{\text{face}}$ and $\mathbf{PA}_{\text{face}}$ are also given each of them calculating the respective error measure only for the class face.

3.2.2 Metrics for vital signs estimation

As metrics for evaluating the vital signs estimates, the commonly employed mean absolute error (**MAE**) and root-mean-square error (**RMSE**) are calculated. This involves computing the difference between estimated value (EST) and ground truth (GT). They are defined as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |EST_i - GT_i| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (EST_i - GT_i)^2} \tag{14}$$

where N denotes the total number of signal windows.

In addition, a further metric is given depending on the respective vital sign.

In case of estimating the HR, the **IEC Accuracy** is additionally applied. It originates from the IEC standard 60601-2-27, which is used for benchmarking medical electrocardiogram devices. Therefore, it is well suited for the given purpose and has already been employed in other works (Rapczynski et al. 2018, 2019). It gives the percentage of correctly determined windows, whereby a window is classified as correctly determined as soon as the error between estimated HR and ground truth is smaller than 10% of the ground truth or smaller than 5 beats per minute (bpm), depending on which of both is higher.

When estimating the RR, the detection rate (**DR**) is used as a further metric indicating the percentage of correctly estimated windows out of the total number of windows. A window is assumed to be correct if the error between estimated value and ground truth is less or equal than 2 brpm. It was similarly introduced by Charlton et al. (2018) and additionally utilized in other studies (Fiedler et al. 2020, 2021).

3.3 Benchmarking

For evaluation, the algorithms will be benchmarked against state-of-the-art procedures to allow high quality conclusions about the obtained results. The methods employed for this benchmark are therefore further explained in this section. First, it starts by describing lightweight architectures for real-time semantic segmentation. Afterwards, ROIs from the area of rPPG estimation are presented. At the end, algorithms for estimating vital signs are explained further by dividing them between the estimation of HR and RR.

3.3.1 Lightweight semantic segmentation architectures

Most of the popular Deep Learning based models for semantic segmentation utilize some kind of encoder-decoder architecture (Minaee et al. 2021). We focus exclusively on lightweight networks, as only these are reasonably applicable for the usage as ROI in rPPG, which requires the segmentation masks to be rendered in real-time for each frame of a video. Details on the training process of the CNNs can be found in Sect. 2.3.

The first end-to-end trainable lightweight network for semantic segmentation without any extra post-processing steps was **ENet** (Paszke et al. 2016). It has been designed by connecting a large number of bottleneck layers in series to reduce the computational cost and is inspired by the ResNet (He et al. 2016a) architectures. **BiSeNet** (Yu et al. 2018) combines the output of a shallow network to encode rich spatial information together with deep network features to provide sufficient receptive field. **LEDNet** (Wang et al. 2019) applies in its encoder special residual blocks with

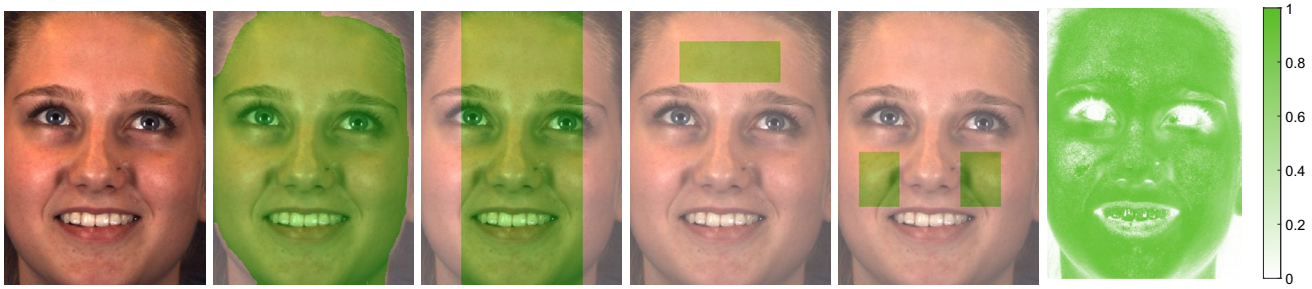


Fig. 6 Example of different ROIs on the MMSE-HR/BP4D+ dataset: (**a**) showing the face crop from the face detection, (**b**) the face segmented by FaSeNet, (**c**) the FaceMid, (**d**) the forehead, (**e**) the cheeks and (**f**) the skin segmentation with corresponding probabilities (see color bar)

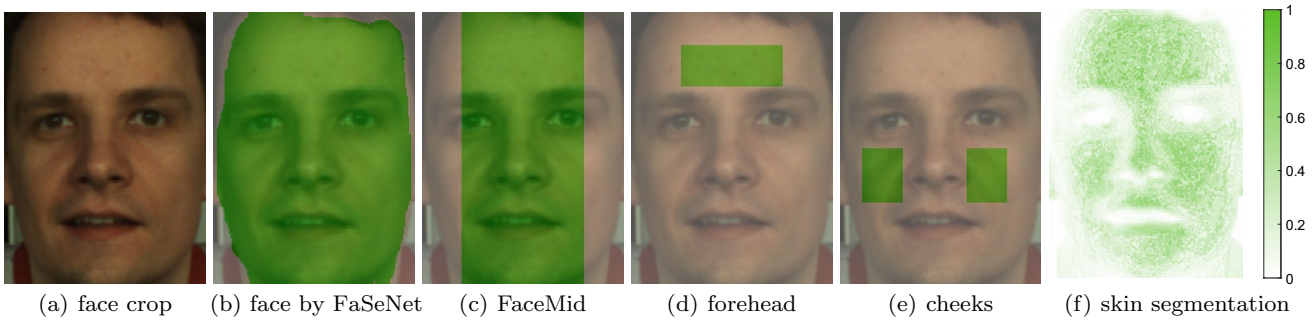


Fig. 7 Example of different ROIs on the PURE dataset: (**a**) showing the face crop from the face detection, (**b**) the face segmented by FaSeNet, (**c**) the FaceMid, (**d**) the forehead, (**e**) the cheeks and (**f**) the skin segmentation with corresponding probabilities (see color bar)

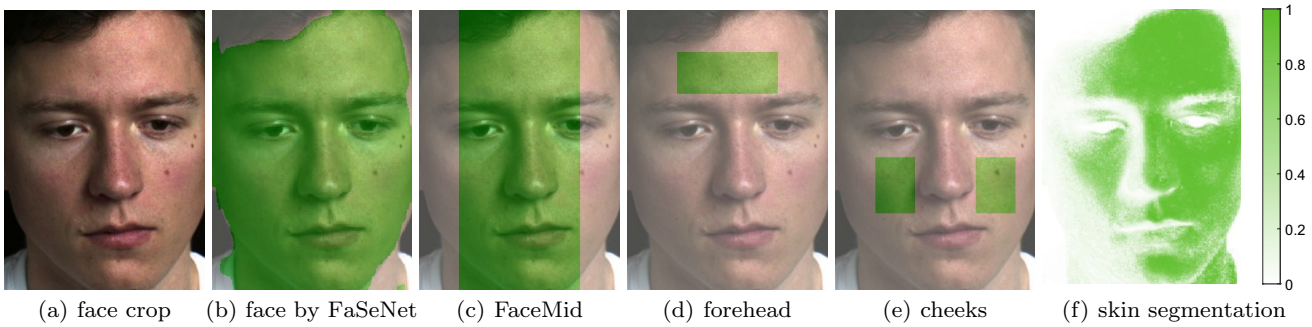


Fig. 8 Example of different ROIs on our own database: (**a**) showing the face crop from the face detection, (**b**) the face segmented by FaSeNet, (**c**) the FaceMid, (**d**) the forehead, (**e**) the cheeks and (**f**) the skin segmentation with corresponding probabilities (see color bar)

channel split and shuffle operations and in its decoder an attention pyramid network. In **DFANet** (Li et al. 2019), subnetwork and sub-stage aggregation cascades are used to join discriminative features while reducing the number of parameters. **Fast-SCNN** (Poudel et al. 2019) introduces a "learning to downsample" module for computing features on multiple resolution branches simultaneously in order to combine spatial detail with deep receptive fields. In **HLNet** (Feng et al. 2020), an information fusion network is proposed which integrates high-dimensional and low-dimensional feature maps in parallel.

3.3.2 Region-of-interests (ROIs)

Selecting the ROI is a crucial step in any vital signs estimation algorithm, as it determines which pixels of an image sequence are used as source for the rPPG signal. If a signal of poor quality is already generated at this early stage in the processing chain, this will lead to bad performance in the downstream estimation of HR and RR. To enable comparison of the proposed CNN-based ROI using FaSeNet, several popular ROIs from existing research were re-implemented. This includes multiple face regions. All ROIs are shown in Figs. 6, 7 and 8.

All ROIs obtain as starting point face crops generated by CNN from the face detection algorithm RetinaFace (Deng et al. 2020) for further processing. Based on these face crops the exact ROI is computed for every frame. In addition, the **face crop** is also used as an individual ROI without further segmentation. Thereby, all pixels within the crop are included in the average calculation for the rPPG signal.

The following ROIs which are based on static geometric regions are provided: The **forehead** is a commonly used region in rPPG (Verkruysse et al. 2008; Sanyal and Nundy 2018; Blöcher et al. 2017). Its width is defined by setting 0.5 of the crop width and its height by calculating 0.3 of the distance between eye corners and bottom of nose. It is placed above the eyebrows. Another popular ROI is **FaceMid** (Poh et al. 2010, 2011; Monkaresi et al. 2017), which takes the full height of the face bounding box but only 0.6 of the centered width. Additionally, the **cheeks** are applied as ROI which was originally introduced by Feng et al. (2015). It is calculated by employing an affine transformation and tracking speeded-up-robust-feature points in the center of the face to estimate its motion.

Furthermore, a pixel-based **skin segmentation** was proposed as ROI by Rapczynski et al. (2018). It uses a lookup table approach which provides for each pixel a relative probability of being skin. This skin probability forms the weighting factor by which the color values of each pixel are included in the average calculation of the rPPG signal. Thus, thresholding and binary masking are avoided. The lookup table for skin segmentation was trained on the ECU (Phung et al. 2005) dataset.

3.3.3 Algorithms for HR estimation

Some prior work algorithms, which are widely used in other research studies, were re-implemented in order to be able to investigate the influence of differing ROIs. For this purpose, the subsequent processing with the estimation of HR from the rPPG signals generated out of these ROIs remains unchanged for analyzing exclusively the photoplethysmographic information present in the signals and their suitability for estimation methods of vital signs.

In the following, the algorithms are briefly explained. But their functional principle is only roughly described, extensive details can be found in the original papers. It should be noted that the ROI applied in the respective paper is not employed, instead each of the ROI approaches presented in Sect. 3.3.2 are executed. The window length is uniformly set to 30 s with a step size of 10 s.

One of the first publications addressing video-based HR estimation was by **Poh et al.** (2011). Thereby, the RGB channels are smoothed and decomposed into independent source signals with the help of an independent component analysis (ICA). Afterwards, filtering and interpolation of the signal with the highest power spectrum peak takes place to derive the inter-beat intervals (IBIs).

Feng et al. (2015) utilize an adaptive green/red color difference (aGRD) operation to calculate the desired rPPG signal. From this signal the HR frequency can be estimated and an adaptive bandpass filter is created to further remove noise and motion artifacts.

Wang et al. (2017a) transform the RGB color channels into the frequency domain and later reconstruct the rPPG signal back to the time domain under consideration of the frequency range by using defined color-channel combinations and underlying weights of their skin model.

A graph-based HR estimation algorithm was presented by **Rapczynski et al.** (2016) which applies an adaptive bandpass on the signal window. Different kind of rPPG signals are used by the authors like the green channel or one of the two signal extraction techniques normalized green (normG) (Stricker et al. 2014) or chrominance-based method (CHROM) (de Haan and Jeanne 2013). Subsequently, smaller signal windows are analyzed and bandpass filtered in a smaller range, whereby the previously determined HR frequency forms the center of this pass band. Out of this, the signal peaks are isolated and all possible connections are plotted in a graph to find out the sequence which minimizes the error based on IBI analysis.

Sanyal and Nundy (2018) transform the pixels of the ROI from RGB into HSV color space. Only the hue channel and pixel values in a predefined range are used to form the rPPG signal. After bandpass filtering the HR is determined in the frequency spectrum.

3.3.4 Algorithms for RR estimation

Similar to the previous subsection, methods for estimating the RR with fixed window length of 60 s and step size of 10 s are also provided:

When determining the RR, **Poh et al.** (2011) additionally analyze the generated IBIs in the Lomb periodogram within the range of human respiration (see Sect. 3.3.3).

Sanyal and Nundy (2018) adopt appropriately adjusted bandpass cutoff frequencies corresponding to possible RR frequencies instead of HR as in 3.3.3.

Furthermore, **FuseMod** (Fiedler et al. 2020) was introduced as a method which modulates seven signal parameters influenced by respiration, including one amplitude modulation, three baseline modulations, and three frequency modulations. These signal parameters are derived from the respective rPPG signal being employed, such as the green channel, normG (Stricker et al. 2014) or CHROM (de Haan and Jeanne 2013). The final RR is received by calculating the median of the seven RR estimates afterwards.

FuseModV2 (Fiedler et al. 2021) is an extension of FuseMod (Fiedler et al. 2020) that adds as eighth modularity the interpolation of differences between rise and fall times of the rPPG signal, making the method more robust.

4 Results and discussion

In this section, the experimental results are presented in detail and afterwards discussed. First, the segmentation performance of the individual networks for the given segmentation task is considered and compared. Subsequently, the speed of the architectures along with their accuracy during inference is examined in order to analyze the execution efficiency. Thereafter, the usage of the MCC loss function is compared with commonly applied loss functions. At the end, the application of a wide variety of ROIs in several rPPG algorithms is evaluated for both HR and RR estimation methods.

4.1 Segmentation performance

The segmentation performance of our FaSeNet network was compared with other commonly used state-of-the-art semantic segmentation architectures on both LFW-PL and CelebHair datasets. The results for LFW-PL are shown in Table 1, where training was performed on the two subsets train and validation followed by inference on the test subset. In Table 2 the results for CelebHair are provided as 5-fold cross-validation, since the authors did not specify a database split.

For LFW-PL, our FaSeNet outperforms all other architectures across all performance metrics (see Sect. 3.2.1), the only exception builds PA_{face} where BiSeNet (Yu et al. 2018) performs better by 0.05%. However, for our main measure the $F1_{face}$ score only FaSeNet achieves a value above 97%,

Table 1 Results comparison on LFW-PL test set in % of our FaSeNet and other network architectures trained on LFW-PL trainval

Network	$F1_{face}$	mF1	IoU_{face}	mIoU	fwIoU	PA_{face}	PA	mPA
DFANet (Li et al. 2019)	95.43	84.31	91.61	77.09	87.99	97.43	93.13	84.87
Fast-SCNN (Poudel et al. 2019)	96.15	86.53	92.78	79.57	89.59	96.13	93.83	89.74
ENet (Paszke et al. 2016)	96.17	85.89	92.90	79.03	89.32	97.43	93.96	85.91
LEDNet (Wang et al. 2019)	96.41	86.92	93.26	80.26	89.99	97.53	94.29	87.68
HLNet (Feng et al. 2020)	96.43	86.93	93.30	80.20	90.03	97.50	94.27	87.62
BiSeNet (Yu et al. 2018)	96.95	88.60	94.22	82.53	91.53	**98.07**	95.29	88.80
FaSeNet (ours)	**97.21**	**89.48**	**94.69**	**83.72**	**92.16**	98.02	**95.65**	**89.88**

Table 2 5-fold cross-validation results comparison on CelebHair in % of our FaSeNet and other network architectures

Network	$F1_{face}$	mF1	IoU_{face}	mIoU	fwIoU	PA_{face}	PA	mPA
DFANet (Li et al. 2019)	95.89	85.55	92.24	77.40	87.46	97.52	92.85	87.13
HLNet (Feng et al. 2020)	96.20	86.17	92.79	78.37	88.29	97.66	93.28	87.41
ENet (Paszke et al. 2016)	96.29	85.70	92.96	77.70	87.92	96.43	92.96	88.42
Fast-SCNN (Poudel et al. 2019)	96.36	86.04	93.08	78.27	88.44	96.92	93.26	89.32
LEDNet (Wang et al. 2019)	96.38	87.14	93.13	79.60	88.85	97.92	93.63	89.34
BiSeNet (Yu et al. 2018)	97.02	88.16	94.28	80.98	90.07	97.87	94.40	90.20
FaSeNet (ours)	**97.06**	**88.79**	**94.34**	**81.89**	**90.50**	**98.04**	**94.68**	**91.21**

the others are all below. In addition, a value of 89.48% is achieved for mF1, which is about 0.9% above BiSeNet (Yu et al. 2018) and the other networks perform significantly worse here with values below 87%. These observations are also reflected in the different IoU parameters. DFANet (Li et al. 2019) is clearly the worst performer, while Fast-SCNN (Poudel et al. 2019), ENet (Paszke et al. 2016), LEDNet (Wang et al. 2019), and HLNet (Feng et al. 2020) are on an equal level.

These findings are also confirmed by the results on CelebHair. DFANet (Li et al. 2019) performs worst across all metrics, while FaSeNet ranks best for each of them. BiSeNet (Yu et al. 2018) is again in second place with a difference of approximately 0.1−1.0% depending on the particular measure.

Overall, it is clear to see the dominance of our FaSeNet network over the others when considering the segmentation results for both only the face and all three classes. Incorrect pixel classifications occur mainly between the two classes hair and background, which becomes evident through the weaker results for the averaged metrics (mF1, mIoU and mPA) compared to those focusing on the face ($F1_{face}$, IoU_{face} and PA_{face}). False positive and false negative segmented pixels for the face class are primarily found at the edges of the face. Other error-prone cases are male beards, although a full beard is usually not a problem for FaSeNet. Correct classification is particularly difficult for less dense beards, where skin remains slightly visible through the structure of the beard.

4.2 Inference time comparison

Besides evaluating solely the segmentation performance, this was also considered in relation to the inference speed, since it is of special importance when using CNNs as ROI in rPPG algorithms that they must be able to capture the pulse-synchronous color changes of the skin in real-time. For this purpose, the number of floating point operations (FLOPs) and the total number of model parameters (Params) are listed for all architectures in addition to the speed and some performance measures in Table 3. The test environment was identical to the one used for training, running a NVIDIA RTX 2080 Ti along with the TensorFlow framework (see Sect. 2.3).

Fast-SCNN (Poudel et al. 2019) and HLNet (Feng et al. 2020) have the fastest inference speeds with 33.42 fps, while FaSeNet with the best segmentation results is in the mid-table with 27.06 fps still ensuring real-time capability. DFANet (Li et al. 2019) has the fewest FLOPs, ENet (Paszke et al. 2016) the fewest Params. Nevertheless, it can be seen that the meaningfulness of FLOPS and Params in terms of inference efficiency is very limited, since these two networks are the ones with the lowest speed.

Looking at the final execution efficiency, it can be stated that FaSeNet ensures a good trade-off between speed and best segmentation. If higher frame rates above 30 fps are required, Fast-SCNN (Poudel et al. 2019) and HLNet (Feng et al. 2020) can represent good alternatives.

Table 3 Results comparison in terms of execution efficiency of our FaSeNet and other network architectures

Network	Speed (fps)	FLOPs (G)	Params (M)	$F1_{face}$ (%)		mF1 (%)	
				LFW-PL	CelebHair	LFW-PL	CelebHair
DFANet	21.80	**0.08**	0.43	95.43	95.89	84.31	85.55
ENet	23.20	0.94	**0.37**	96.17	96.29	85.89	85.70
LEDNet	23.85	3.28	2.30	96.41	96.38	86.92	87.14
FaSeNet	27.06	11.27	24.23	**97.21**	**97.06**	**89.48**	**88.79**
BiSeNet	28.48	2.70	26.38	96.95	97.02	88.60	88.16
Fast-SCNN	**33.42**	0.41	1.62	96.15	96.36	86.53	86.04
HLNet	**33.42**	0.94	1.23	96.43	96.20	86.93	86.17

Table 4 Results comparison on LFW-PL test set in % of our FaSeNet using different loss functions trained on LFW-PL trainval

Loss function	$F1_{face}$	mF1	IoU_{face}	mIoU	fwIoU	PA_{face}	PA	mPA
Dice	96.96	89.01	94.24	83.05	91.69	98.01	95.38	89.16
Jaccard	97.12	89.28	94.52	83.40	91.99	**98.10**	95.55	89.46
Cross-entropy	97.16	88.52	94.61	82.53	91.79	97.80	95.05	87.81
MCC	**97.21**	**89.48**	**94.69**	**83.72**	**92.16**	98.02	**95.65**	**89.88**

4.3 Loss function comparison

Using the FaSeNet architecture, the application of different loss functions and their impact on the overall performance of the network was investigated. The results after training on the LFW-PL trainval set and testing on the LFW-PL test set can be found in Table 4. These were listed exemplarily, but are also consistent with other network architectures as well as on the CelebHair dataset. The standard loss functions Dice, Jaccard, and Cross-Entropy, which are certainly the most commonly used ones in the vast majority of papers in the field of CNN-based segmentation, were employed for comparison purposes.

The MCC (see Sect. 2.2) outperforms all other losses across all metrics except for PA_{face} where Jaccard is better by 0.08%. However, MCC is superior to Jaccard in the remaining scores by 0.1–0.4%. Jaccard is in turn superior to Dice in all metrics, while Cross-Entropy is only able to outperform Jaccard in $F1_{face}$ and IoU_{face}, but does not reach the values of MCC. It is interesting to observe that the MCC loss even performs better on the metrics F1 and IoU than the respective version optimized exactly for this parameter (with Dice and Jaccard loss respectively). This is a striking illustration of how well the MCC is suited as a training loss and how powerful the inclusion of true negative samples into the metric is affecting the results.

Altogether, the great potential of the MCC as a loss function in multi-class segmentation becomes apparent. This is especially valid in case of datasets which are characterized by an imbalance in the occurrence probabilities of individual classes, as it is the case in our work.

4.4 Vital signs estimation performance

We have evaluated the ROI approaches in extensive experiments with a variety of algorithms for vital signs estimation. See Sect. 3.3 for details about the ROI and vital signs estimation approaches. A distinction was made between methods for measuring HR and RR. For each of the two modalities, experiments were conducted with two different databases to obtain generally valid results. Therefore, in addition to MAE and RMSE, the IEC Accuracies for the different ROI approaches have been calculated for each HR method, they can be found for PURE in Tables 5, 6 and 7 and for MMSE-HR in Tables 8, 9 and 10. The same was done for the MAE, RMSE and DR results for RR estimation on BP4D+ and our own database, which are shown in Tables 11, 12 and 13 and 14, 15 and 16 respectively. In addition, for each ROI the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ across all rPPG algorithms were computed and reported.

For PURE the findings are obvious: FaSeNet achieves the lowest MAE and RMSE as well as the highest IEC Accuracy over all algorithms, which reaches its MAE minimum with 2.82 bpm and DR maximum with 95.77% for Sanyal and Nundy (2018). For RMSE, Wang et al. (2017a) performed best with 9.92 bpm compared to second best Sanyal and Nundy (2018) with 10.15 bpm. The mean values for the three metrics MAE, RMSE and IEC Accuracy are also significantly better than that of skin segmentation with a plus of approximately 1.4 bpm, 2.8 bpm and 4.2% respectively. But skin segmentation still performs better than the other remaining ones. This dominance of FaSeNet can be explained by the significantly poorer lighting conditions in PURE, which cause difficulties for the skin segmentation, as it can be seen in Fig. 7f. Regarding the constantly high standard deviations, it should be noted that the algorithm of (Poh et al. 2011 struggles with this dataset regardless of the ROI, which however was not investigated in more detail.

On MMSE-HR, the algorithms of Rapczynski et al. (2016) with normG, Wang et al. (2017a), and Sanyal and Nundy (2018) perform best for the IEC Accuracy, while the one of Feng et al. (2015) performs worst. But it is interesting to observe that Poh et al. (2011) is the top performer for metric RMSE with 5.54 bpm and in second place for MAE with 3.16 bpm. This points to the importance of evaluating various error measures in order to be able to draw generally valid conclusions. In particular, Poh et al. (2011) appears to report smaller errors relative to the ground truth for correctly as well as incorrectly estimated windows, which is not taken into account by the IEC Accuracy. Thus, the method demonstrates a high degree of robustness. There are variations in the ROIs that have the highest IEC Accuracy for each method: FaSeNet is ahead three times, skin segmentation twice, and face crop and forehead once each. A similar behavior is observed for MAE and RMSE, but in this case the skin segmentation is ahead only once, the number of top positions for the others remains unchanged. Considering the mean IEC Accuracy, FaSeNet and skin segmentation are also in front with values above 88%, while the others do not even reach 86%. The same can be observed for MAE and RMSE, where only FaSeNet and skin segmentation are in the mean under 5 bpm and 11 bpm, respectively. FaSeNet is on average 0.14% better for the IEC Accuracy than skin segmentation, but the individual results deviate more strongly from one another which is reflected in a 1.32% higher standard deviation. This is also consistent with the results of MAE and RMSE. The only exception is the standard deviation for RMSE, which is 0.16 bpm lower for FaSeNet than for skin segmentation. This suggests that more outliers occur when using skin segmentation. Overall, the results for MMSE-HR indicate that the ROIs FaSeNet, skin segmentation, and face crop are close to each other. We explain this effect by the fact that strong movements of the subjects occur in MMSE-HR, which leads to steady changes of the head pose. As a result, skin pixels disappear from the visible area of the image, while others reappear. ROIs that include a large number of

Table 5 HR estimation results comparison on the PURE dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as MAE in bpm

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	24.92	**16.48**	22.84	22.35	25.88	16.93
Feng et al. (2015)	aGRD	6.21	**3.58**	5.05	6.51	6.36	3.67
Wang et al. (2017a, b)	RGB	6.22	**2.92**	7.03	4.62	6.72	4.66
Rapczynski et al. (2016)	CHROM	10.48	**10.24**	10.64	10.35	10.43	13.63
	normG	8.86	**4.79**	7.62	5.12	8.69	7.47
Sanyal and Nundy (2018)	hue	3.64	**2.82**	3.09	2.95	3.72	2.86
μ		10.06	**6.81**	9.38	8.65	10.30	8.20
σ		7.66	**5.49**	7.07	7.16	7.96	5.79

Table 6 HR estimation results comparison on the PURE dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as RMSE in bpm

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	29.15	**22.47**	28.24	28.27	29.93	22.48
Feng et al. (2015)	aGRD	13.41	**10.37**	12.58	18.77	13.09	10.65
Wang et al. (2017a, b)	RGB	18.94	**9.92**	21.68	13.46	20.41	15.50
Rapczynski et al. (2016)	CHROM	22.10	**21.63**	22.54	24.60	22.02	26.18
	normG	19.97	**16.15**	18.25	18.12	20.56	20.86
Sanyal and Nundy (2018)	hue	11.75	**10.15**	11.22	10.60	11.95	10.44
μ		19.22	**15.12**	19.09	18.97	19.66	17.89
σ		6.28	**5.86**	6.44	6.62	6.56	6.51

Table 7 HR estimation results comparison on the PURE dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as IEC Accuracy in %

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	17.69	**41.92**	20.77	29.62	15.77	37.31
Feng et al. (2015)	aGRD	76.15	**92.31**	86.15	86.54	81.12	91.54
Wang et al. (2017a, b)	RGB	88.46	**92.31**	86.92	86.15	84.73	91.15
Rapczynski et al. (2016)	CHROM	73.46	**79.62**	71.54	78.08	73.41	67.31
	normG	79.62	**91.92**	81.54	91.15	80.65	85.77
Sanyal and Nundy (2018)	hue	91.92	**95.77**	95.00	94.62	91.88	95.38
μ		71.22	**82.31**	73.65	77.69	71.26	78.08
σ		27.17	**20.55**	27.02	24.20	27.84	22.31

Table 8 HR estimation results comparison on the MMSE-HR dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as MAE in bpm

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	4.31	**3.16**	4.08	4.14	4.52	3.31
Feng et al. (2015)	aGRD	**8.20**	9.11	9.80	12.90	12.67	8.44
Wang et al. (2017a, b)	RGB	4.37	3.53	5.40	**3.46**	4.80	3.70
Rapczynski et al. (2016)	CHROM	9.67	**5.72**	12.83	7.66	10.04	6.76
	normG	2.81	2.89	3.42	4.69	4.38	**2.59**
Sanyal and Nundy (2018)	hue	5.31	**3.68**	5.16	3.90	5.89	3.86
μ		5.78	**4.68**	6.78	6.13	7.05	4.78
σ		2.61	2.39	3.71	3.64	3.48	**2.29**

Table 9 HR estimation results comparison on the MMSE-HR dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as RMSE in bpm

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	7.95	**5.54**	7.33	8.41	8.81	6.59
Feng et al. (2015)	aGRD	**16.01**	17.08	18.34	23.36	22.56	17.40
Wang et al. (2017a, b)	RGB	10.58	8.34	13.01	**7.66**	11.97	8.60
Rapczynski et al. (2016)	CHROM	19.83	**14.26**	22.64	17.23	20.66	15.44
	normG	9.63	9.80	11.29	12.59	12.45	**8.45**
Sanyal and Nundy (2018)	hue	12.02	**9.03**	11.79	9.43	12.72	9.14
μ		12.67	**10.68**	14.07	13.11	14.86	10.94
σ		4.44	**4.22**	5.50	6.13	5.45	4.38

Table 10 HR estimation results comparison on the MMSE-HR dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as IEC Accuracy in %

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	85.03	**89.82**	85.63	86.83	84.07	89.22
Feng et al. (2015)	aGRD	76.05	75.45	72.46	69.46	70.14	**79.64**
Wang et al. (2017a, b)	RGB	89.82	92.81	89.22	**93.41**	89.60	91.62
Rapczynski et al. (2016)	CHROM	77.84	**86.83**	68.26	83.23	75.42	83.83
	normG	**94.61**	94.51	92.81	88.62	89.77	**94.61**
Sanyal and Nundy (2018)	hue	88.02	**92.53**	88.62	92.22	87.48	92.22
μ		85.23	**88.66**	82.83	85.63	82.75	88.52
σ		7.15	7.01	10.02	8.74	8.16	**5.69**

Table 11 RR estimation results comparison on the BP4D+ dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as MAE in brpm

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	5.41	**4.86**	5.45	5.04	5.62	5.27
Sanyal and Nundy (2018)	hue	**7.08**	7.48	7.31	7.73	7.55	7.47
FuseMod	CHROM	2.34	**1.97**	2.30	2.29	2.49	2.00
	normG	2.41	**2.16**	2.59	2.68	2.73	2.20
FuseModV2	CHROM	2.26	**1.94**	2.25	2.32	2.30	1.99
	normG	2.35	**2.11**	2.49	2.73	2.68	2.18
μ		3.64	**3.42**	3.73	3.80	3.90	3.52
σ		**2.09**	2.29	2.14	2.18	2.18	2.32

Table 12 RR estimation results comparison on the BP4D+ dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as RMSE in brpm

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	6.87	**6.42**	6.94	6.52	6.95	6.88
Sanyal and Nundy (2018)	hue	**8.05**	8.40	8.33	8.54	8.52	8.38
FuseMod	CHROM	3.40	**2.95**	3.37	3.26	3.48	2.97
	normG	3.58	**3.09**	3.64	3.75	3.81	3.12
FuseModV2	CHROM	3.31	**2.90**	3.25	3.29	3.27	2.91
	normG	3.50	**3.03**	3.55	3.85	3.69	3.12
μ		4.79	**4.47**	4.85	4.87	4.95	4.56
σ		**2.11**	2.37	2.21	2.17	2.22	2.42

Table 13 RR estimation results comparison on the BP4D+ dataset for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as DR in %

Algorithm	rPPG signal	ROIs					
		face crop	face by FaSeNet	FaceMid	forehead	cheeks	skin segmentation
Poh et al. (2011)	ICA	33.25	**40.93**	34.03	36.26	33.01	37.01
Sanyal and Nundy (2018)	hue	**16.75**	14.39	15.74	10.47	12.75	14.58
FuseMod	CHROM	66.98	**72.02**	67.36	65.23	64.83	71.78
	normG	65.09	**69.11**	62.04	60.93	60.56	68.22
FuseModV2	CHROM	68.40	**73.23**	68.98	64.49	65.14	73.08
	normG	66.51	**69.45**	63.19	59.25	60.74	68.60
μ		52.83	**56.52**	51.89	49.44	49.51	55.55
σ		22.20	23.93	21.85	21.89	**21.68**	24.18

Table 14 RR estimation results comparison on our own database for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as MAE in brpm

Algorithm	rPPG signal	ROIs					
		face crop	face by FaSeNet	FaceMid	forehead	cheeks	skin segmentation
Poh et al. (2011)	ICA	3.90	**2.73**	3.78	3.42	3.94	2.81
Sanyal and Nundy (2018)	hue	2.07	**1.65**	1.92	2.49	2.50	1.81
FuseMod	CHROM	1.42	**0.83**	1.18	1.66	1.71	0.86
	normG	1.65	**1.05**	1.88	1.97	2.00	1.06
FuseModV2	CHROM	1.40	**0.75**	1.18	1.64	1.67	0.79
	normG	1.55	**0.79**	1.68	1.97	1.99	0.80
μ		2.00	**1.30**	1.94	2.19	2.30	1.36
σ		0.96	0.78	0.96	**0.68**	0.86	0.81

Table 15 RR estimation results comparison on our own database for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as RMSE in brpm

Algorithm	rPPG signal	ROIs					
		face crop	face by FaSeNet	FaceMid	forehead	cheeks	skin segmentation
Poh et al. (2011)	ICA	5.50	**4.42**	5.30	5.08	5.41	4.46
Sanyal and Nundy (2018)	hue	3.50	**2.95**	3.36	3.89	3.48	3.20
FuseMod	CHROM	3.05	**2.10**	2.64	3.28	3.30	2.20
	normG	3.26	**2.46**	3.48	3.69	3.78	2.49
FuseModV2	CHROM	2.88	**2.08**	2.50	3.13	3.22	2.10
	normG	3.01	**2.31**	3.10	3.55	3.67	**2.31**
μ		3.53	**2.72**	3.40	3.77	3.81	2.79
σ		0.99	0.89	1.01	**0.70**	0.81	0.91

Table 16 RR estimation results comparison on our own database for several algorithms with our FaSeNet (trained on LFW-PL trainval and Celeb-Hair) and other state-of-the-art ROIs as DR in %

Algorithm	rPPG signal	ROIs					
		Face crop	Face by FaSeNet	FaceMid	Forehead	Cheeks	Skin segmentation
Poh et al. (2011)	ICA	48.05	**62.40**	47.11	53.08	46.13	59.25
Sanyal and Nundy (2018)	hue	74.10	**80.19**	76.91	68.41	67.99	78.63
FuseMod	CHROM	82.37	**89.86**	84.40	76.45	75.37	87.68
	normG	77.07	**85.02**	72.54	74.02	72.81	84.87
FuseModV2	CHROM	82.22	**90.87**	83.78	77.20	76.24	90.09
	normG	78.32	**90.11**	75.20	72.71	75.55	90.09
μ		73.69	**83.08**	73.32	70.31	69.02	81.77
σ		12.95	10.92	13.68	**9.00**	11.61	11.83

pixels can compensate for these effects better than smaller ones during averaging the rPPG signal.

The BP4D+ dataset is a very challenging database for RR estimation, as it is characterized by strong movements of the subjects because they had to complete certain tasks while being recorded. This difficulty particularly affects the method of Sanyal and Nundy (2018) which is not capable of compensating for these strong interferences at all. Its highest MAE, RMSE and DR is achieved for the face crop at just 7.08 brpm, 8.05 brpm and 16.75%, respectively. But this is the only algorithm that does not perform best together with FaSeNet. For the others, the familiar trend recurs with the skin segmentation ranking slightly behind. The best MAE, RMSE and DR is shown for FuseModV2 (Fiedler et al. 2021) with CHROM using FaSeNet at 1.94 brpm, 2.90 brpm and 73.23%, while skin segmentation is 0.05 brpm, 0.01 brpm and 0.15% worse, respectively. Looking at the DR mean values, FaSeNet is at the top by 56.52% with a difference of about 1% compared to skin segmentation. The other ROIs are clearly behind with face crop next with 52.83%. This is also in accordance with the results for MAE and RMSE.

Also for our own database, FaSeNet proves its dominance and performs best for each RR estimation method and all three metrics. Skin segmentation again follows in second place. FuseModV2 (Fiedler et al. 2021) is again at the top with DRs of 90.87% for CHROM and 90.11% for normG. MAE and RMSE are 0.75 brpm and 2.08 brpm for CHROM and 0.79 brpm and 2.31 brpm for normG, respectively. It is noticeable that for RMSE, FuseMod (Fiedler et al. 2020) with CHROM outperforms FuseModV2 (Fiedler et al. 2021) with normG by scoring a value of 2.10 brpm. But for the other two metrics, this is not the case. The skin segmentation in combination with FuseModV2 (Fiedler et al. 2021) also reaches DRs above 90% and is only closely behind FaSeNet. For the RMSE measure and the rPPG signal normG, FuseModV2 (Fiedler et al. 2021) even yields the exact same value of 2.31 brpm for both FaSeNet and skin segmentation. All other ROIs are clearly beaten and achieve a maximum of slightly more than 73% on average for DR. This is around 10% behind FaSeNet and around 8% behind skin segmentation. Furthermore, their MAE and RMSE scores exceed 1.90 brpm and 3.00 brpm, respectively, while they are approximately 1.30 brpm and 2.70 brpm for FaSeNet and skin segmentation. The gap between the two top performers face by FaSeNet and skin segmentation can possibly be explained by the fact that there was a slight shadow cast into the subjects' faces, causing the skin segmentation not to classify some facial pixels correctly (see Fig. 8f). This illustrates the benefit of CNN-based segmentation, which incorporates semantic information and is resistant to illumination variations.

All in all, it can be stated that classification models including the whole facial area (face by FaSeNet and skin segmentation) perform significantly better as ROI than static geometric regions of the face (FaceMid, forehead and cheeks). This was clearly shown over all three metrics for both HR and RR estimation as well as across all four employed databases. Our developed FaSeNet and the skin segmentation deliver the rPPG signals with the best photoplethysmographic information. In particular, in scenarios with suboptimal ambient illumination, CNN-based face segmentation can demonstrate its benefits. In video recordings with good and constant illumination, both achieve a similar level of performance for the vital signs estimation, but still with slight advantages for FaSeNet. Thus, it was finally demonstrated that the proposed FaSeNet is highly effective for segmenting faces and subsequently generating rPPG signals out of these frames. Hence, a ROI was created that is superior to other approaches in camera-based vital signs estimation.

Our FaSeNet represents a contribution to the current state of the art, although it is not capable of solving all presently existing problems in the field of vital signs estimation. The main challenge that still remains are head movements of subjects. Such movements cause skin pixels to disappear from the visible area of the image while others reappear, leading to artifacts in the rPPG signal. This in turn results in higher errors for the estimated vital signs. Therefore, the compensation of these motion artifacts is of immense importance for the further progress of the research in this area. Possible approaches for future work could be to integrate the head pose information into the rPPG signal generation process in order to learn how to compensate for it. For this purpose, the temporal head pose changes may be analyzed in a downstream network, which is supposed to use this knowledge for suppressing the motion artifacts in the rPPG signal.

5 Conclusion

Our newly proposed FaSeNet achieves better results in face segmentation than other commonly used architectures from the state of the art. This finding could be proven on the two datasets LFW-PL and CelebHair. In addition, it achieves a fast inference speed in real-time resulting in an overall high execution efficiency with respect to the segmentation performance. The newly applied loss function based on the MCC was able to provide a better fitting of the CNN weights for the face segmentation task than the loss functions Dice, Jaccard, and Cross-Entropy, which are commonly used in the field of multi-class segmentation. In particular, the MCC is well suited as a loss function for datasets that are

characterized by an imbalanced class distribution through its inclusion of true negative samples into the metric. In an extensive evaluation with a variety of algorithms for vital signs estimation, it was proven that by utilizing our FaSeNet as ROI better results could be achieved for both HR and RR estimation compared to static facial regions or a lookup table based skin segmentation as ROI. This enabled to overcome the limitations of those ROI approaches in the presence of partial occlusions by interfering pixels, caused e.g. by hair or headgear, in case of skin tone variations, as well as changing, too weak or too strong ambient illumination scenarios by employing CNN-based segmentation. For future work, it may be attempted to incorporate head pose information into the ROI processing to compensate for motion artifacts. In conclusion, by applying our FaSeNet, a new ROI superior to previous approaches was created for vital signs estimation from video images.

Acknowledgements Open Access funding provided by Projekt DEAL. This work was supported by the German Research Foundation (DFG) under grants AL 638/13-1 and AL 638/14-1.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Becker C et al (2017) Camera-based measurement of respiratory rates is reliable. Eur J Emerg Med. https://doi.org/10.1097/mej.0000000000000476

Blöcher T et al (2017) An online PPGI approach for camera based heart rate monitoring using beat-to-beat detection". In: IEEE sensors applications symposium (SAS). https://doi.org/10.1109/sas.2017.7894052

Borza D, Ileni T, Darabant A (2018) A deep learning approach to hair segmentation and color extraction from facial images. In: International conference on advanced concepts for intelligent vision systems, pp 438–449

Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLOS ONE. https://doi.org/10.1371/journal.pone.0177678

Castaneda D et al (2018) A review on wearable photoplethysmography sensors and their potential future applications in health care. Int J Biosens Bioelectron. https://doi.org/10.15406/ijbsbe.2018.04.00125

Charlton PH et al (2018) Breathing rate estimation from the electrocardiogram and photoplethysmogram: a review. IEEE Rev Biomed Eng 11:2–20. https://doi.org/10.1109/rbme.2017.2763681

Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. https://doi.org/10.1186/s12864-019-6413-7

Chicco D, Tötsch N, Jurman G (2021) The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. https://doi.org/10.1186/s13040-021-00244-z

de Haan G, Jeanne V (2013) Robust pulse rate from chrominance-based rppg. IEEE Trans Biomed Eng 60(10):2878–2886. https://doi.org/10.1109/tbme.2013.2266196

Deng J et al (2009) ImageNet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2009.5206848

Deng J et al (2020) Retinaface: single-shot multilevel face localisation in the wild. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00525

Elliott M, Coventry A (2012) Critical care: the eight vital signs of patient monitoring. Br J Nurs 21(10):621–625. https://doi.org/10.12968/bjon.2012.21.10.621

Fel L, Malik M (1994) Heart rate variability. Card Pac Electrophysiol. https://doi.org/10.1007/978-94-011-0872-0_6

Feng L et al (2015) Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. IEEE Trans Circuits Syst Video Technol 25(5):879–891. https://doi.org/10.1109/tcsvt.2014.2364415

Feng X, Gao X, Luo L (2020) HLNet: a unified framework for real-time segmentation and facial skin tones evaluation. Symmetry. https://doi.org/10.3390/sym12111812

Fiedler M-A, Rapczynski M, Al-Hamadi A (2020) Fusion-based approach for respiratory rate recognition from facial video images. IEEE Access. https://doi.org/10.1109/access.2020.3008687

Fiedler M-A, Rapczynski M, Al-Hamadi A (2021) Facial video-based respiratory rate recognition interpolating pulsatile PPG rise and fall times. In: IEEE 18th international symposium on biomedical imaging (ISBI). https://doi.org/10.1109/isbi48211.2021.9434132

Fouad RM, Omer OA, Aly MH (2019) Optimizing remote photoplethysmography using adaptive skin segmentation for real-time heart rate monitoring. IEEE Access 7:76513–76528. https://doi.org/10.1109/access.2019.2922304

Garcia-Garcia A et al (2018) A survey on Deep Learning techniques for image and video semantic segmentation. Appl Soft Comput 70:41–65. https://doi.org/10.1016/j.asoc.2018.05.018

He K et al (2016a) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2016.90

He K et al (2016b) Identity mappings in deep residual networks. In: European conference on computer vision (ECCV), pp 630–645. https://doi.org/10.1007/978-3-319-46493-0_38

Howard AG et al (2017) MobileNets: efficient convolutional neural networks for mobile vision applications

Huang N et al (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. https://doi.org/10.1371/journal.pgen.1001154

Jadon S (2020) A survey of loss functions for semantic segmentation. In: IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). https://doi.org/10.1109/cibcb48159.2020.9277638

Kae A et al (2013) Augmenting CRFs with Boltzmann machine shape priors for image labeling. In: IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2013.263

Lever J, Krzywinski M, Altman N (2016) Classification evaluation. Nat Methods 13(8):603–604. https://doi.org/10.1038/nmeth.3945

Li H et al (2019) DFANet: deep feature aggregation for real-time semantic segmentation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2019.00975

Minaee S et al (2021) Image segmentation using deep learning: a survey. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/tpami.2021.3059968

Monkaresi H et al (2017) Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affect Comput 8(1):15–28. https://doi.org/10.1109/taffc.2016.2515084

Moraes J et al (2018) Advances in photopletysmography signal analysis for biomedical applications. Sensors 18(6):1894. https://doi.org/10.3390/s18061894

Nilsson LM et al (2007) Combined photoplethysmographic monitoring of respiration rate and pulse: a comparison between different measurement sites in spontaneously breathing subjects. Acta Anaesthesiologica Scandinavica. https://doi.org/10.1111/j.1399-6576.2007.01375.x

Nisar H et al (2016) Contactless heart rate monitor for multiple persons in a video. In: IEEE international conference on consumer electronics-Taiwan (ICCE-TW). https://doi.org/10.1109/icce-tw.2016.7520988

Paszke A et al (2016) ENet: a deep neural network architecture for real-time semantic segmentation

Phung SL, Bouzerdoum A, Chai D (2005) Skin segmentation using color pixel classification: analysis and comparison. IEEE Trans Pattern Anal Mach Intell 27(1):148–154. https://doi.org/10.1109/tpami.2005.17

Poh M-Z, McDuff DJ, Picard RW (2010) Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt Express 18(10):10762. https://doi.org/10.1364/oe.18.010762

Poh M-Z, McDuff DJ, Picard RW (2011) Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans Biomed Eng 58(1):7–11. https://doi.org/10.1109/tbme.2010.2086456

Poudel R, Liwicki S, Cipolla R (2019) Fast-SCNN: fast semantic segmentation network

Rapczynski M, Werner P, Al-Hamadi (2016) Continuous low latency heart rate estimation from painful faces in real time. In: 23rd international conference on pattern recognition (ICPR). https://doi.org/10.1109/icpr.2016.7899794

Rapczynski M, Werner P, Saxen F et al (2018) How the region of interest impacts contact free heart rate estimation algorithms. In: 25th IEEE international conference on image processing (ICIP). https://doi.org/10.1109/icip.2018.8451846

Rapczynski M, Werner P, Al-Hamadi A (2019) Effects of video encoding on camera-based heart rate estimation. IEEE Trans Biomed Eng 66(12):3360–3370. https://doi.org/10.1109/tbme.2019.2904326

Sanyal S, Nundy KK (2018) Algorithms for monitoring heart rate and respiratory rate from the video of a user's face. IEEE J Transl Eng Health Med. https://doi.org/10.1109/jtehm.2018.2818687

Song J et al (2006) Prediction of cis/trans isomerization in proteins using psi-blast profiles and secondary structure information. BMC Bioinform. https://doi.org/10.1186/1471-2105-7-124

Stricker R, Müller S, Gross H-M (2014) Non-contact video-based pulse rate measurement on a mobile service robot. In: 23rd IEEE international symposium on robot and human interactive communication. https://doi.org/10.1109/roman.2014.6926392

Ulku I, Akagündüz E (2022) A survey on deep learning-based architectures for semantic segmentation on 2D images. Appl Artif Intell. https://doi.org/10.1080/08839514.2022.2032924

Verkruysse W, Svaasand LO, Nelson JS (2008) Remote plethysmographic imaging using ambient light. Opt Express 16(26):21434. https://doi.org/10.1364/oe.16.021434

Wang CY et al (2015) imDC: an ensemble learning method for imbalanced classification with miRNA data. Genet Mol Res 14(1):123–133. https://doi.org/10.4238/2015.january.15.15

Wang W, den Brinker AC et al (2017a) Robust heart rate from fitness videos. Physiol Meas 38(6):1023–1044. https://doi.org/10.1088/1361-6579/aa6d02

Wang W, Stuijk S, de Haan G (2017b) Living-skin classification via remote-ppg. IEEE Trans Biomed Eng 64(12):2781–2792. https://doi.org/10.1109/tbme.2017.2676160

Wang C, Pun T, Chanel G (2018) A comparative survey of methods for remote heart rate detection from frontal face videos. Front Bioeng Biotechnol. https://doi.org/10.3389/fbioe.2018.00033

Wang Y et al (2019) LEDNet: a lightweight encoder-decoder network for real-time semantic segmentation. In: IEEE international conference on image processing (ICIP). https://doi.org/10.1109/icip.2019.8803154

Yu C et al (2018) BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: European conference on computer vision, pp 334–349. https://doi.org/10.1007/978-3-030-01261-8_20

Zhang Z et al (2016) Multimodal spontaneous emotion corpus for human behavior analysis. In: IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr.2016.374

Zhao L, Reisman S, Findley T (1994) Derivation of respiration from electrocardiogram during heart rate variability studies. Comput Cardiol. https://doi.org/10.1109/cic.1994.470251

Zhu Q (2020) On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognit Lett 136:71–80. https://doi.org/10.1016/j.patrec.2020.03.030