Weakly supervised human skin segmentation using guidance attention mechanisms

Kooshan Hashemifard[1] · Pau Climent-Perez[1] · Francisco Florez-Revuelta[1]

© The Author(s) 2023

Abstract
Human skin segmentation is a crucial task in computer vision and biometric systems, yet it poses several challenges such as variability in skin colour, pose, and illumination. This paper presents a robust data-driven skin segmentation method for a single image that addresses these challenges through the integration of contextual information and efficient network design. In addition to robustness and accuracy, the integration into real-time systems requires a careful balance between computational power, speed, and performance. The proposed method incorporates two attention modules, Body Attention and Skin Attention, that utilize contextual information to improve segmentation results. These modules draw attention to the desired areas, focusing on the body boundaries and skin pixels, respectively. Additionally, an efficient network architecture is employed in the encoder part to minimize computational power while retaining high performance. To handle the issue of noisy labels in skin datasets, the proposed method uses a weakly supervised training strategy, relying on the Skin Attention module. The results of this study demonstrate that the proposed method is comparable to, or outperforms, state-of-the-art methods on benchmark datasets.

Keywords Skin segmentation · Attention mechanism · Weakly supervised training · Deep neural networks

1 Introduction

Skin detection, also known as skin segmentation, is the process of separating skin pixels or regions in an image from non-skin pixels such as background or covered body pixels [1]. This technique has a wide range of applications, including human biometric analysis, medical

✉ Kooshan Hashemifard
k.hashemifard@ua.es

Pau Climent-Perez
pau.climent@ua.es

Francisco Florez-Revuelta
francisco.florez@ua.es

1 Department of Computing Technology, University of Alicante, St. San Vicente del Raspeig, San Vicente del Raspeig 03690, Alicante, Spain

image analysis, autonomous driving, and the beauty industry. Additionally, it is also important in other applications such as content retrieval, robotics, sign language recognition, and human tracking [2]. Furthermore, skin segmentation is often the first step in tasks related to nudity and appearance detection [3]. The use of cameras in video-based applications for Ambient Assisted Living (AAL) and remote monitoring of patients raises privacy concerns, particularly related to unwanted nudity. To address these concerns, a promising solution is privacy by context [4], which adapts privacy levels based on various factors, including the level of nudity. In these applications, the accurate detection of skin areas is critical for ensuring privacy.

However, skin segmentation is a challenging task as it faces many difficulties including variations in illumination conditions, different skin tones, camera variations, makeup, ageing, and backgrounds with similar colours. The success of skin segmentation algorithms depends on their ability to overcome these challenges and accurately identify skin pixels in an image.

Prior to the advent of deep learning, most methods relied on the distribution of skin colours to identify fixed or adaptive ranges of skin pixel values in different colour spaces or the training of a classifier according to each colour space [5]. However, these methods have several drawbacks, such as poor generalization and high dependence on the training data. As a result, they tend to have poor performance in complex background situations and high false positive detection rates, making them ill-suited for real-world scenarios.

With the rise of deep learning, skin segmentation accuracy has improved drastically. The introduction of Fully Convolutional Networks (FCN) for semantic segmentation by Shelhamer et al. [6], as well as increases in computation power, have boosted both accuracy and speed in the segmentation task. Additionally, other methods such as UNet [7] and DeepLab [8] were proposed, capable of segmenting multiple objects with high precision. These methods have also been adapted for human skin segmentation [5, 9]. Despite the advances in accuracy achieved by these deep learning-based methods, there is still room for improvement. In particular, they are highly dependent on large datasets and tend to perform poorly on small or noisy skin segmentation datasets [10].

In this paper, we propose a novel skin segmentation method that leverages the advances in human body segmentation to improve accuracy. In recent years, human body segmentation has improved significantly, with reliable open source models such as DensePose [11] and Mask R-CNN [12] being published, which can provide different body parts regions in a given image. These regions can serve as contextual information for detecting skin pixels. Instead of approaching the problem of skin segmentation from scratch, our method utilizes this auxiliary information to increase accuracy with lighter models. Our approach involves the use of two separate attention modules in the decoder, namely the Body Attention module and the Skin Attention module. The Body Attention module utilizes the output of the body mask to make the network focus on the body boundary. In the Skin Attention module, we make the assumption that the face and hands areas consist mostly of skin pixels. Therefore, we calculate the affinity of these pixels and body pixels embeddings to produce a skin attention mask, which provides extra guidance to the network for the target class. By integrating these attention modules, our method can effectively segment skin pixels by utilizing the information from human body segmentation and facial features.

Our contribution to the field of skin segmentation includes:

- A lightweight model that improves skin segmentation performance without greatly increasing model size.
- Incorporating contextual information in the skin detection process to enhance performance.

- Modifying commonly used attention mechanisms to better suit the skin segmentation task.
- Utilizing individual skin colour as a cue for segmentation to address variations in skin tone.
- A weakly supervised training strategy that utilizes the proposed attention module for training on noisy datasets.

The above contributions provide a method that balances the trade-off between accuracy, resources and computational cost which make it feasible for real-world applications.

The remaining structure of this paper is as follows: Section 2 provides a brief overview of current methods for human skin detection. Section 3 describes the proposed method for skin segmentation, including the architecture and the details of attention modules. Section 4 presents experiments and results, as well as an ablation study. Section 5 introduces a weakly supervised training strategy to improve results. Finally, in Section 6, conclusions of the work and suggestions for future research are discussed.

2 Related work

Human skin segmentation is a challenging task in computer vision and machine learning, and has been widely studied in the literature. Similar to many other recognition and segmentation tasks, human skin segmentation can be broadly categorized into two groups: traditional methods, and deep learning methods.

Traditional skin detection methods have primarily focused on skin colour characteristics. For example, thresholding methods based on the intensity of skin colour have been proposed in [13, 14]. These methods typically define a set of rules and conditions for each colour channel, and check if a given pixel satisfies these conditions to decide if it belongs to the skin class or not. These rules are established either based on trial and error [15–17] or by using thresholding algorithms [18]. However, these methods have a high false detection rate and are strongly dependent on the training dataset and its conditions.

To address the limitations of traditional skin detection methods, various researchers have proposed more robust methods by dynamically updating rules using cues in the image. For example, Shifa et al. [19] proposed a hybrid method with Combined Threshold-rules that adapts the threshold ranges by detecting the sampling skin tone and analysing the colour histogram and distribution. Probabilistic methods were also proposed as a way of detecting skin by evaluating the general distribution of skin colour. These methods include using parametric and non-parametric techniques such as histograms, look-up tables (LUT), Naive Bayes and Gaussian distributions. For example, Gomez et al. [20] used 3D histograms to propose the conditions for skin probability, utilizing a simplified version of probability theory. Nanni et al. [21] proposed the use of histograms with multiple LUTs in different colour spaces, which resulted in a reduction of the false detection rate. However, these methods, despite their speed, do not take into account the relation between adjacent pixels and the spatial information which are crucial for skin detection [22]. Bayesian classifiers have also been used for skin detection, but they require a large training set to achieve high accuracy. For instance, [23–25] used the Bayes rule to calculate the conditional probability density function of a pixel from normalized histograms and adjust threshold values. Additionally, some works [26–28] have modelled skin colour distribution using Gaussian mixtures, 2D Gaussian Probability Density Function (PDF) were used to model skin colour in different colour spaces like YUV, HSV

and RGB. These methods typically require the use of expectation maximization (EM) for parameter optimization [29]. Supervised learning and binary classification methods, such as Support Vector Machines (SVMs) and Multi-Layer Perceptrons (MLPs), have also been employed in skin detection. These methods train a classifier on extracted colour-texture features, such as Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), FAST and ORB, derived from blobs of an image [30, 31]. For example, Han et al. [32] used SVM for binary pixel-wise classification. Another example is the combination of SVM and Gaussian Mixture Model (GMM) proposed in [33].

Region-based skin detection is another approach that has been used in skin detection. These techniques aim to make use of spatial information and connections in an image to identify regions with similar features, without building a skin colour model. Region growing is one such technique where the segmentation process starts with selected seed pixels. The neighbouring pixels are compared to the seed pixel and if they have similar properties, they are added to the seed pixel's region, which then grows. This process is repeated for new adjacent pixels until no neighbouring pixels satisfy the conditions and the growth of the region stops. Variations of this method have been proposed in [34–36] where different measurement techniques have been proposed to measure the similarity between adjacent pixels for assigning them to a region, such as Euclidean distance, colour distance map or probability measuring.

With the advent of deep learning, newer ideas have been proposed for human skin segmentation, many of which draw inspiration from the broader research area of semantic segmentation or a combination of various deep learning concepts. Early work treated the problem as a classification task by dividing an image into smaller patches and using deep networks to perform binary classification between skin and non-skin classes [37–39]. For example, Lei et al. [38] proposed a patch-based skin segmentation method using stacked autoencoders to extract discriminative features from the blobs in an image, but this approach is not efficient in terms of time and resources required and it does not take into account relations between patches and contextual information.

Fully Convolutional Networks (FCNs) [6] have become the main approach for various segmentation tasks, including skin segmentation, due to their ability to reduce the number of parameters through an encoder-decoder architecture. The encoder extracts features and performs down-sampling, while the decoder up-samples the features to the original input size. For example, Kim et al. [37] evaluated some of the prevalent FCN architectures for skin segmentation. Chang-Hsian et al. [40] used pre-trained ResNet50 and transfer learning. Roy et al. [41] combined conditional adversarial training approach [42] along with U-Net [7]. Zuo et al. [9] used RNN layers as they claim that CNN are not sufficient to model the relationship between adjacent pixels. He et al. [5] proposed a semi-supervised method which takes advantage of other similar datasets and used dual-task fully convolutional network which shares the encoder and two separate decoders for detecting both body and skin parts in a U-Net shape auto-encoder. Arsalan et al. [43] proposed an end-to-end semantic segmentation network with an outer residual skip connection to transfer the edge information from early layers to end layers.

A successful technique in semantic segmentation is the use of self-attention modules for modelling long-range dependencies within an image. This technique was originally proposed in [44] for machine translation but has since been widely used in various tasks such as [45, 46]. In computer vision, self-attention modules were introduced in [47] to extract global dependencies of inputs for better image generation. Since then, various kinds of self-attention

mechanisms have been proposed, ranging from simple lightweight methods such as [48] to more complex methods like [49] that capture dependencies in different dimensions.

3 Method

In this section, we present a novel method for efficient skin segmentation. We begin by describing the task-oriented attention modules that form a key component of our approach. Next, we outline the general architecture of our network, which is based on a fully convolutional network (FCN) with encoder-decoder architecture. Finally, we discuss the efficiency of our approach, specifically in terms of the reduction of parameters in the encoder.

3.1 Attention mechanisms

The attention modules are designed to help the network focus on specific areas of an image that are relevant to the task at hand, such as skin pixels in this case, without adding a large amount of redundant parameters. The two attention mechanisms used in this work are:

- Body Attention: This is used to emphasize the body area in a given image, as the skin pixels are only present within the body boundaries. This module is used in an earlier stage of the decoder, and its purpose is to guide the network to focus on the body area and extract body-related information from the encoded data.
- Skin Attention: This module is designed to compare the embeddings of pre-defined skin areas, such as the face and hands, to all other body embeddings. The purpose of this module is to provide auxiliary guidance to refine the output. It is implemented in a later stage of the decoder with a higher resolution, which leads to finer boundaries and reduces noise.

By using these two attention mechanisms, the network is able to perform more efficiently by focusing on the given task and reach higher accuracy using fewer parameters. In the following, each module is explained in detail.

3.1.1 Body attention module

The Body Attention module aims to focus feature extraction on the body area. It is based on the Convolutional Block Attention Module (CBAM) [48] method for adaptive feature refinement, but with slight modifications to make it better suited for the task of skin segmentation.

CBAM produces attention maps along the channel and spatial dimensions sequentially. The module starts by generating an attention map for the channel dimension by using a fully connected layer and a sigmoid activation function to weigh the importance of each channel. Then, it generates an attention map for the spatial dimension by using max-pooling and average-pooling to compute the importance of each spatial location. These two attention maps are then multiplied element-wise to produce a final attention map. This map is then used to weight the feature maps and refine the feature extraction.

This modified version of CBAM is used in the Body Attention module to emphasize the body area in a given image, and guide the network to focus on the body area and extract body-related information from the encoded data. This helps to improve the performance of the network by focusing on the most important areas of the image.

For a given block input tensor $F \in \mathbb{R}^{C \times H \times W}$, a 1D channel attention $M_c \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$ are inferred [48]. According to [50] each

channel of feature tensor acts as an object detector, therefore channel attention tries to find meaningful objects in an image.

For the channel attention submodule, we followed the CBAM approach described in detail by Woo et al. [48]. Specifically, we use the following steps:

1. The input tensor is max-pooled and average-pooled simultaneously and squeezed spatially.
2. The descriptor vectors are then fed to a shared network, which is a multi-layer perceptron (MLP) with one hidden layer. The size of the hidden layer is reduced from the input by a ratio of ∇.
3. The outputs of the shared MLP are then merged by summation and a sigmoid function is applied to the result.
4. Finally, this channel attention vector is broadcasted to the input tensor F.

The spatial attention submodule focuses on the area where the desired object (skin) appears. It uses a modified version of the CBAM approach to compute the spatial attention. In CBAM, the spatial attention is computed by applying average-pooling and max-pooling operations along the channel axis and concatenating them to generate an efficient feature descriptor. This descriptor is then convolved by a standard convolution layer. In our approach, we make use of the knowledge that skin is present within the body boundaries. We extract a body mask by using a pre-trained network and concatenate it to the original concatenation before the convolution layer. This forces the attention module to put emphasis on the area of interest (human body) and extract more related attention maps from the input. The Sigmoid function is again applied to the final result. It has been empirically confirmed in our experiments that by adding the body mask in the concatenation operation, the network can extract a more related attention map from the input, which results in an improvement in accuracy of the whole network.

This refined and task-oriented CBAM submodule can be summarized as follows:

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$

and

$$M_s(F') = \sigma(Conv^{7 \times 7}(Concat[AvgPool(F'); MaxPool(F'); BodyMask(I)])) \tag{4}$$

The Sigmoid function is denoted by σ, F' is the input tensor to the spatial attention and I is the input image. The output 2D map is multiplied to all input channels of F'. Finally, using a skip connection, the attention module output F'' is summed with the module input tensor F.

3.1.2 Skin attention module

Skin detection in images can be a difficult task, especially in unconstrained situations where skin areas can be exposed on different body parts with different shapes and under varying body poses. Conventional CNNs, which are designed to find objects and shapes for a given class, may not be completely effective or appropriate for this task. In addition, Fully Convolutional Networks with convolution operations have local receptive fields and fail to capture long-range contextual information. This means that the features extracted from the same class in

different parts of the image may vary in representation, leading to a performance drop due to intra-class inconsistency. Furthermore, to account for the variety of human skin colours, the datasets for the deep learning model have to be large and inclusive of all skin types. To overcome these challenges, a skin attention module inspired by the work of Fu et al. [49] is introduced in this study. The proposed method is able to take advantage of a wider range of contextual information and draw a global context over local features to improve feature representations for better pixel-level prediction and noise reduction. The mechanism is based on the assumption that face and hand areas of a given human image mostly consist of skin pixels.

In this proposed method, we begin by using a pre-trained body segmentation network to infer the body part masks from the input image, I. In the attention module, an intermediate input feature tensor $T \in \mathbb{R}^{C \times H \times W}$ is duplicated and then fed to a CNN layer, followed by batch normalization to obtain K, $Q \in \mathbb{R}^{C \times H \times W}$. Then, by using the body part masks, the binary mask of face and hand area is multiplied by broadcasting to K, and the full body binary mask is element-wise multiplied to Q. Both outcomes are then reshaped to $\mathbb{R}^{C \times N}$ where $N = H \times W$. The energy matrix $E \in \mathbb{R}^{N \times N}$ is calculated by taking the matrix multiplication of the transpose of Q and K.

By utilizing the body part masks, we then element-wise multiply the binary mask of the face and hand area with K and the full body binary mask with Q. Both matrices are reshaped to C×N. The energy matrix, E, with dimensions N×N, is calculated by taking the matrix multiplication of the transpose of Q and K.

Each non-zero column in the reshaped matrices K and Q represents the feature vector of face pixel and body pixel, respectively. Therefore, each element in E represents the inner product or similarity measure between a pixel in the face and hands areas and a pixel in the body. By calculating the weighted average of E with respect to the number of face and hands pixels in I, the average similarity between the embedding of a pixel in body area and all the pixels in hands and face areas is obtained. The closer they are, the output of the inner product is higher.

The resulting matrix, called the similarity matrix $S \in \mathbb{R}^{N \times 1}$, is further processed by applying the tanh function for normalization and limiting the values below 1. It is then reshaped to $1 \times H \times W$, which has the same dimensions as the input tensor. This resulting matrix, called the attention map A, illustrates the correlation between a body pixel and all supposed skin pixels in the hands and face of a person. Since the body of each person is compared to their corresponding face and hands, this method works well with a variety of skin colours. Additionally, all body embeddings are compared to the face regardless of location, body part, or the shape of that body part. This allows for a broader range of contextual information to be involved in the process, rather than relying solely on local features from the CNN.

To enable the network to adjust the effect of this attention map, a trainable weight parameter, ω, is applied to it. This allows the network to decide the importance of this attention map and use it accordingly. The final attention map is then added to the input tensor, T, to obtain the final result.

Empirical results demonstrate that the trainable weight parameter, ω, tends to increase during the training process. This suggests that the attention map, A, contains useful information for the final prediction. The proposed attention module, as a whole, is depicted in Fig. 1. This highlights the effectiveness of the proposed method in capturing the correlation between pixels in the body, face and hands areas, and utilizing it to improve the final prediction.
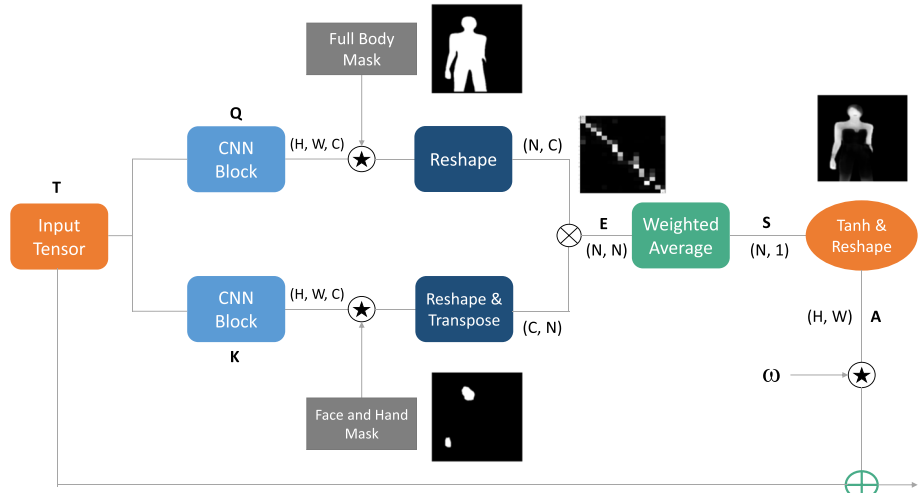
Fig. 1 Skin Attention Module. ⊛ denotes element-wise multiplication and × is matrix multiplication

3.2 Network architecture

Semantic segmentation has seen significant advancements in recent years, with several methods achieving high accuracy on benchmark datasets. However, these methods often rely on heavy backbone networks, making them less suitable for real-time applications with limited computation resources. To address this issue, researchers have proposed lightweight networks that can achieve performance comparable to that of high-quality networks while consuming less computational resources. Examples of such networks include ENet [51], DFANet [52], and LEDNet [53], which employ techniques such as depthwise separable convolutions (DwConv2D) [54], feature aggregation subnetworks, and channel split and shuffle.

The proposed model for semantic segmentation consists of two channels. The primary channel of the network is an asymmetric encoder-decoder architecture that incorporates auxiliary submodules in the pipeline. It is inspired by HLNet [55] and HRNet [56] and employs best practices in segmentation modules.

The encoder consists of CNNs and DwConv2D, bottlenecks and interaction modules. The input image, with a resolution of 256×256, is fed to a CNN block with 32 filters, followed by two DwConv2D blocks with 64 filters and a stride of 2. This fast downsampling process ensures low-level feature sharing [57]. Each block has 3×3 kernels, followed by batch normalization and ReLU activation. To preserve details and constrain the number of parameters, the maximum downsampling rate is set to 1/8. The output, with a resolution of 32×32, is then fed to the information interaction module proposed in [55]. This module consists of three parallel inverted residual blocks using different filter sizes and strides. Each block captures information with different resolutions and feature map sizes, learning multiscale information representation. This process increases the network's ability to segment small objects and shapes while preserving more details. Subsequently, the information from high to low resolutions is combined together by concatenation. For more details, please refer to HLNet [55].

In addition to the primary channel, the proposed model also includes an auxiliary channel that is pre-trained for body segmentation. Given an image, this model produces a whole

body mask, as well as separate masks for the face and hands. These masks are then resized and fed to the decoder to be used in attention mechanisms. These attention mechanisms (see Section 3.1) aim to make use of the information provided by the auxiliary channel to improve the overall performance of the semantic segmentation.

The decoder of the proposed model consists of three bilinear upsampling layers, each followed by a convolution block, to construct an original size image from the 32×32 feature map in the bottleneck. To make use of the information provided by the auxiliary channel, a body attention module is applied after the first upsampling layer and a skin attention module is applied after the second one. Additionally, skip connections are added to propagate the error to the early feature extraction layers.

Finally, a Sigmoid layer is applied to perform binary classification between skin and non-skin pixels. In order to optimize the network, given the class imbalance between skin and non-skin pixels in most images, where the background is larger than the area covered by the person, a combination of Dice loss and Binary Focal loss is used between the output mask and the ground truth. This has been shown to be more effective for segmenting small objects [58]. The overall network architecture is illustrated in Fig. 2.

4 Experimental results

4.1 Datasets and implementation details

Obtaining data for skin segmentation is a common challenge in this field. One approach is to collect bespoke datasets, but there are also several public datasets available for this task which can be useful for the evaluation and comparison of methods, even though they may not follow the same protocol (i.e., some considered eyebrows and lips as skin and
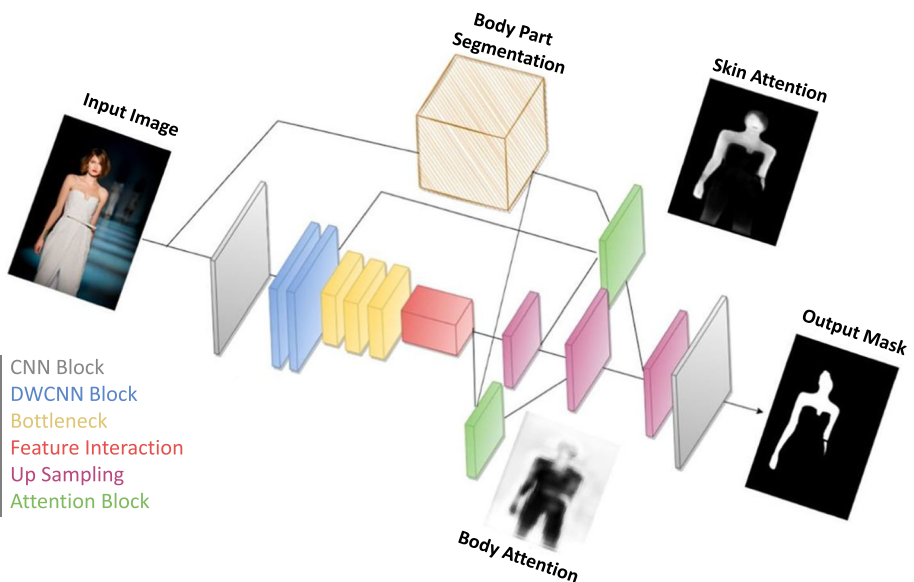

Fig. 2 Proposed skin segmentation network

some excluded them) or contain noise. To train the proposed method, the visuAAL Skin Segmentation dataset [10] (VSS) is used. The VSS dataset contains 46,775 high-quality images divided into a training set with 45,623 images and a validation set with 1,152 images. The skin labels were automatically extracted with an algorithm. Additionally, 230 images were manually segmented for evaluation purposes, which are used as the test set in order to report the performance. Some samples of the VSS dataset are shown in Fig. 3. Another dataset incorporated in this work is the Pratheepan face Dataset [59]. Pratheepan is a small dataset containing 78 images with precise annotation and is mostly used as a benchmark. The obtained results are evaluated on this dataset in order to make a comparison with existing methods. Both datasets are publicly available. In this work, the DensePose method was used for body, face and hands area detection. DensePose is a method that can provide precise body masks. To incorporate this information in our network, the parameters of the DensePose network were frozen during the training process.

The proposed method was implemented using TensorFlow and the training was run on two GPUs, one NVIDIA GeForce GTX 1080Ti and one NVIDIA GeForce GTX 2080Ti. The network was trained for 30 epochs using an initial learning rate of 0.001, with a decay rate of 0.96, and the ADAM optimization algorithm. The whole training process took about 5 hours.

4.2 Results

The results of the proposed method were compared with other important semantic segmentation methods on the VSS dataset in Table 1. All of these networks were trained and evaluated on the same data. As shown in the table, the proposed method outperformed all other methods with a considerably lower number of parameters. One important thing to note is the improvement in results when comparing HLNet to our method. Despite their similarities in terms of the backbone network architecture and the amount of parameters employed, the

Fig. 3 Samples from VSS dataset

Table 1 Semantic segmentation results on VSS dataset

Method	Precision	Recall	F1-score	CDR	DSC	Param.
SegNet [60]	80.71%	80.12%	80.29%	97.75%	78.82%	30M
UNet [7]	82.66%	85.34%	83.83%	98.01%	82.38%	30M
Basic FCN [61]	70.34%	84.88%	76.80%	97.11%	74.40%	10M
DSNet [62]	85.80%	85.08%	85.40%	98.35%	84.14%	8M
HLNet [55]	76.50%	79.86%	78.01%	97.43%	76.08%	1.2M
Proposed method	88.30%	85.91%	86.96%	98.45%	86.44%	1M
Base network	76.20%	78.22%	77.19%	97.24%	75.88%	1M
With body attention	78.11%	82.08%	80.04%	97.81%	78.92%	1M
With skin attention	86.62%	83.11%	84.82%	98.18%	83.33%	1M

proposed method shows significant improvement in all metrics for skin segmentation. This demonstrates the effectiveness of the proposed attention mechanisms. An additional study to compare the efficiency of different methods has been performed. The results are illustrated in Table 2 and prove the proposed method is either the best or second only to HLNet but with considerable improvement in the performance metrics.

4.3 Ablation study

Furthermore, to validate the effect of each module, each of them was evaluated separately. The results are presented in Table 1, and it is shown that the Skin Attention played a more important role compared to the Body Attention module, yet adding Body Attention still improved the results from the base network. Figure 4 illustrates some of the segmentation results. One significant observation in the proposed method is the impact of the Body Attention and Skin Attention modules on Recall and Precision. The Body Attention module has a greater effect on Recall compared to Precision, while the Skin Attention improves Precision to a greater extent than recall. This can be explained by the Body Attention's emphasis on the whole body area, resulting in the inclusion of all body pixels and thus, a higher Recall. Conversely, the Skin Attention focuses on excluding non-skin pixels from the final mask, leading to improved Precision. Furthermore, the ω parameter which adjusts the amount of effect and participation of the Attention map in the whole network is set to 1.3 at the end of the training process.

Table 2 Efficiency evaluation

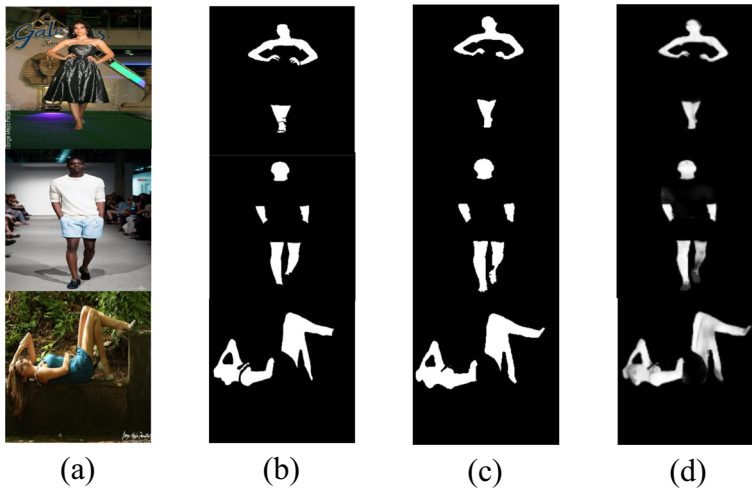| Method	Model size	Training time per epoch	GPU memory required*	Inference time	Param.
SegNet	260MB	30min	430MB	33ms	30M
UNet	120MB	30min	350MB	30ms	30M
Basic FCN	60MB	25min	270MB	30ms	10M
DSNet	120MB	20min	270MB	50ms	8M
HLNet	10MB	10min	70MB	25ms	1.2M
Proposed method	5MB	15min	120MB	30ms	1M

*: For a batch size of one sample

Fig. 4 The proposed method segmentation results on VSS dataset. (a) Original image, (b) ground truth, (c) model output, (d) attention map

5 Improvement using weakly supervised training for noise reduction

Deep learning methods have shown better performance compared to previous approaches for semantic segmentation, but they require a large amount of training data to perform adequately. The annotation process for segmentation is costly, very demanding and labor-intensive. Most of the datasets for skin segmentation are either small or suffering from low-quality images, but the main problem is the labelling noise, as it is expensive to annotate such a large amount of images accurately.

To make use of these big but noisy datasets, a method is proposed to modify the ground truth labels during a recursive training process. This is addressed as a weak supervision task in which, although the labels for a desired class exist (skin pixels), they may be annotated wrongly. In other words, every pixel with a skin class label can either belong to skin or non-skin class, but the pixels labelled as non-skin are considered to be correct. This assumption is made after studying available skin datasets. In this work, the VSS dataset is used, which includes a large training set produced automatically and may contain some noise, and a small fully supervised validation set. The problem of training directly over this dataset is that for a number of epochs, the improvement over validation and training sets accuracy can be seen. However, as the number of epochs increases, the network tends to overfit on the noisy training data, and this causes a drop in validation accuracy. To address this problem, the training data is utilized to the point that it is useful for the objective, and then the noise is modified to redirect the network parameters to segment the desired areas.

This approach allows the network to make use of the large amount of training data while addressing the problem of noisy labels in the dataset. The modified training set is then used to continue training the network for a fixed number of rounds. With this method, the network learns the general segmentation task in the warm-up step and then is able to improve its performance by using the Attention map to correct the noisy labels in the training set in each modification step. The threshold used to relabel the pixels is increased after each modification step, allowing the network to focus on pixels that are more likely to belong to the skin class

while ignoring the noisy labels. Additionally, pixels outside the body boundary area are not relabelled, providing an additional constraint on the modification process.

The Attention map is used before applying the weight parameter ω in the skin attention module. As the Attention map is computed as the product of the tanh function, the values are limited to 1, and can be considered as a skin probability map regarding face and hand pixels. This probability map is then multiplied with the weight parameter ω to adjust the effect and participation of the Attention map in the overall network.

The approach of using the Attention map to modify the ground truth labels during the training process is inspired by previous work [63, 64]. However, two modifications have been made. First, instead of using the output after each round, the Attention map is used. This provides stronger supervision as it uses the auxiliary information provided by the body segmentation module. Second, instead of searching in the rectangular bounding box, the search is limited to the body area (i.e., the only place skin pixels can exist). This reduces the search space and increases the precision of the relabelling process. Additionally, the spatial continuity condition is skipped in this approach. This is because, depending on the garment types, skin exposure can be found in any shape and size over the body.

The labelling procedures for a given pixel in position (i, j) can be written as:

$$L_{(i,j)}^{new} = \begin{cases} 1 & \text{if } L_{(i,j)}^{prev} \otimes P(S \mid F, H)_{(i,j)} > t \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

in which L^{new} will be the new label after this training round, L^{prev} is the last round training ground truths, $P(S \mid F, H)$ is the probability map of the skin class given hand and face and t is the threshold value.
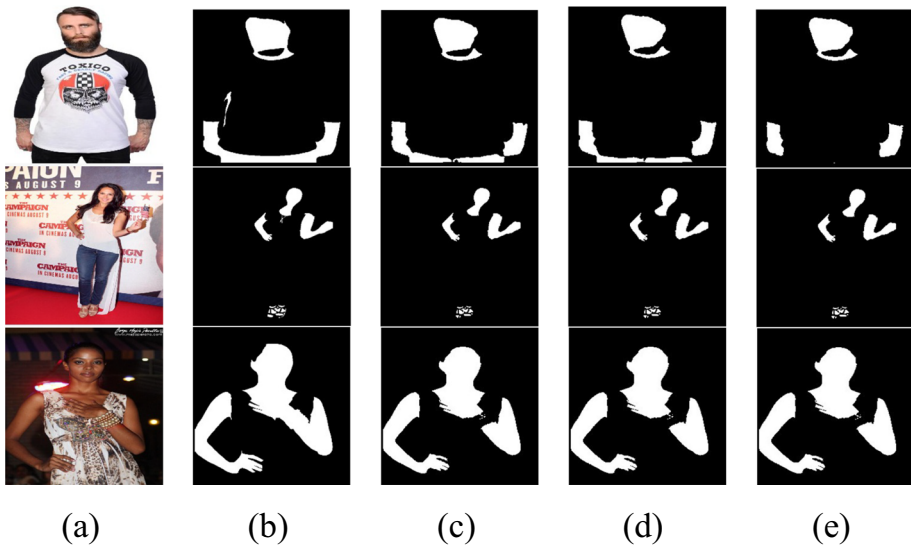

Fig. 5 Recursive training effect on the noisy labels. (a) Original image, (b) noisy labels, (c) after 1st modification round, (d) after 2nd modification round, (e) after 4th modification round

Table 3 Comparing the results of the proposed method on the VSS dataset before and after recursive training

Method	Precision	Recall	F1-score	CDR	DSC
Base network	76.20%	78.22%	77.19%	97.24%	75.88%
Proposed method	88.30%	85.91%	86.96%	98.45%	86.44%
Recursive training	90.59%	87.92%	89.17%	98.61%	88.91%

The improvement in performance is measured using the validation set after each round of the modification step. This process is continued until a performance drop is observed in the validation set. This indicates that the threshold for relabelling pixels has been increased so much that skin pixels are being discarded from the new ground truth labels. It is important to note that this procedure may not remove the noise from all images in the dataset equally or to the same extent, but it corrects the problematic and faulty labels that cause the most errors. As a result, the overall quality of the segmentation improves. In this way, it makes the performance higher on the denoised validation data. For the final step of training, the ground truth labels produced in the epoch previous to the performance drop are considered as the final training set. The training is continued over these labels until the best results are achieved. Depending on the level of noise in each image, the number of modification epochs needed to denoise might be different. As illustrated in Fig. 5, for a small modification such as removing a bracelet or necklace, even one epoch is enough, however, large mislabelling requires more rounds of this recursive training procedure.

In this experiment, the number of training rounds n used for the warm-up step was 2, and the threshold value t for modification started at 0.2 with an increasing step size of 0.05 after each round. The optimum threshold reached at the end was 0.35 and the normal training process started with the dataset generated by this round. In Table 3, the improvement of the results compared with the direct training approach is shown. It can be seen that the proposed method outperforms the direct training approach in terms of overall segmentation performance. This demonstrates the effectiveness of the proposed method in addressing the problem of labelling noise in the training dataset and improving the segmentation results.

In addition, in order to make a comparison between our method and other skin segmentation methods, the evaluation, over the Pratheepan dataset, of the proposed method after the recursive training steps was performed and is illustrated in Table 4. As it can be seen, our method is performing better than classic approaches and produces very similar results to the state-of-the-art method. Considering the model sizes, this can be considered as a higher improvement in efficiency. Some samples of the results on the Pratheepan dataset are illustrated in Fig. 6. This comparison further validates the effectiveness of the proposed method in addressing the problem of labelling noise and improving the overall segmentation performance, while also being more efficient than other methods.

Table 4 Evaluation of human skin segmentation methods on the Pratheepan dataset

Method	Precision	Recall	IoU
Thresholding [65]	65.31%	89.58%	60.20%
GMM [24]	62.36%	91.5%	60.46%
UNet [7]	89.55%	87.87%	85.50%
SOTA [5]	92.58%	87.51%	87.90%
Our method	89.24%	90.95%	89.80%

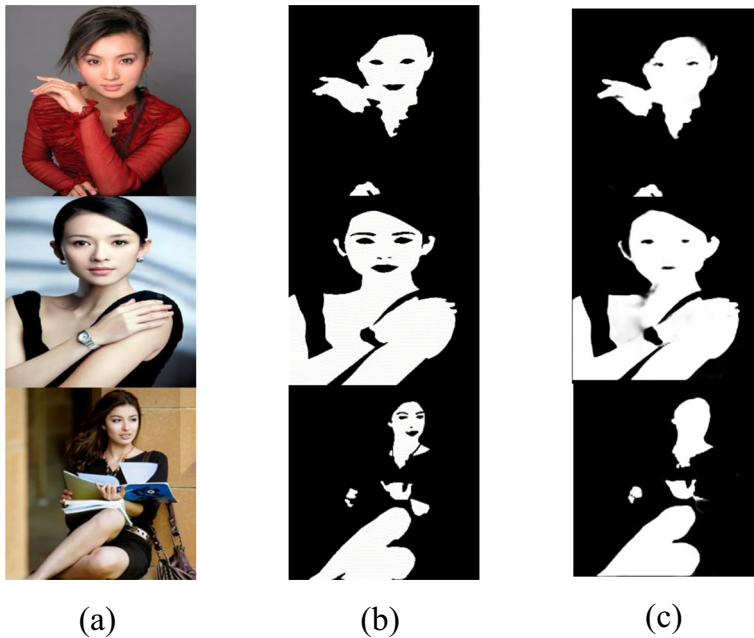(a)   (b)   (c)

Fig. 6 Segmentation results on the Pratheepan benchmark dataset. (a) Original image, (b) ground truth, (c) model output

6 Conclusion

In conclusion, a lightweight, efficient and robust model for human skin segmentation is proposed in this paper. By utilizing prior knowledge and contextual information, the proposed method addresses some of the main challenges in human skin detection, such as variations in skin colour and real-time performance. Additionally, a weakly supervised training strategy is proposed using the attention module to make large datasets with possible annotation errors more usable. The results show that the proposed method outperforms other existing methods in terms of accuracy and efficiency. Furthermore, the method is able to handle unseen skin characteristics and colours. However, to further improve the model, future enhancements such as reducing the memory requirements for calculating the skin attention, adding post-processing steps for smoothing the detected regions and refining the output, and keeping consistency may be beneficial.

Author Contributions All authors contributed to the study conception and design. Implementation and validation were performed by Kooshan Hashemifard. The first draft of the manuscript was written by Kooshan Hashemifard and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

🫧 Springer

Declarations

Ethical and informed consent for data used These datasets have been employed following the requirements established in the datasets' repositories. No own additional data has been acquired and processed in this project.

References

1. Shaik KB, Ganesan P, Kalist V, Sathish B, Jenitha JMM (2015) Comparative study of skin color detection and segmentation in hsv and ycbcr color space. Procedia Computer Science 57:41–48
2. Mahmoodi MR, Sayedi SM (2016) A comprehensive survey on human skin detection. International Journal of Image, Graphics and Signal Processing 8(5):1
3. Maidhof C, Hashemifard K, Offermann J, Ziefle M, Florez-Revuelta F (2022) Underneath your clothes: a social and technological perspective on nudity in the context of aal technology. In: Proceedings of the 15th international conference on PErvasive technologies related to assistive environments, pp 439–445
4. Padilla-López JR, Chaaraoui AA, Gu F, Flórez-Revuelta F (2015) Visual privacy by context: proposal and evaluation of a level-based visualisation scheme. Sensors 15(6):12959–12982
5. He Y, Shi J, Wang C, Huang H, Liu J, Li G, Liu R, Wang J (2019) Semi-supervised skin detection by network with mutual guidance. In: Proceedings of the IEEE/CVF international conference on computer Vision, pp 2111–2120
6. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
7. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, pp 234–241. Springer
8. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
9. Zuo H, Fan H, Blasch E, Ling H (2017) Combining convolutional and recurrent neural networks for human skin detection. IEEE Sig Process Lett 24(3):289–293
10. Hashemifard K, Florez-Revuelta F (2022) From garment to skin: the visuaal skin segmentation dataset. In: International conference on image analysis and processing, pp 59–70. Springer
11. Güler RA, Neverova N, Kokkinos I (2018) Densepose: dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7297–7306
12. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
13. Chaves-González JM, Vega-Rodríguez MA, Gómez-Pulido JA, Sánchez-Pérez JM (2010) Detecting skin in face recognition systems: a colour spaces study. Digital Sig Process 20(3):806–823
14. Yang M-H, Kriegman DJ, Ahuja N (2002) Detecting faces in images: a survey. IEEE Trans Pattern Anal Mach Intell 24(1):34–58

15. Gupta A, Chaudhary A (2016) Robust skin segmentation using color space switching. Pattern Recog Image Anal 26(1):61–68
16. Vadakkepat P, Lim P, De Silva LC, Jing L, Ling LL (2008) Multimodal approach to human-face detection and tracking. IEEE Trans Ind Electron 55(3):1385–1393
17. Do H-C, You J-Y, Chien S-I (2007) Skin color detection through estimation and conversion of illuminant color under various illuminations. IEEE Trans Consum Electron 53(3):1103–1108
18. Santos A, Paiva J, Toledo C, Pedrini H (2016). In: Bhattacharyya S, Dutta P, De S, Klepac G (eds) Improved human skin segmentation using fuzzy fusion based on optimized thresholds by genetic algorithms, pp 185–207. Springer, Cham
19. Shifa A, Imtiaz MB, Asghar MN, Fleury M (2020) Skin detection and lightweight encryption for privacy protection in real-time surveillance applications. Image Vis Comput 94:103589
20. Gomez G (2002) On selecting colour components for skin detection. In: Object recognition supported by user interaction for service robots, vol 2, pp 961–964. IEEE
21. Nanni L, Lumini A, Dominio F, Zanuttigh P (2014) Effective and precise face detection based on color and depth data. Appl Comput Inform 10(1–2):1–13
22. Naji S, Jalab HA, Kareem SA (2019) A survey on skin detection in colored images. Artif Intell Rev 52(2):1041–1087
23. Nadian-Ghomsheh A (2016) Pixel-based skin detection based on statistical models. J Telecommun Electron Comput Eng 8(5):7–14
24. Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. Int J Comput Vis 46:81–96
25. Sigal L, Sclaroff S, Athitsos V (2004) Skin color-based video segmentation under time-varying illumination. IEEE Trans Pattern Anal Mach Intell 26(7):862–877
26. Caetano TS, Olabarriaga SD, Barone DAC (2002) Performance evaluation of single and multiple-gaussian models for skin color modeling. In: Proceedings. XV Brazilian symposium on computer graphics and image processing, pp 275–282. IEEE
27. Liu Z, Yang J, Peng NS (2005) An efficient face segmentation algorithm based on binary partition tree. Signal Process Image Commun 20(4):295–314
28. Shih FY, Cheng S, Chuang C-F, Wang PS (2008) Extracting faces and facial features from color images. Int J Pattern Recognit Artif Intell 22(03):515–534
29. Moon TK (1996) The expectation-maximization algorithm. IEEE Signal Proc Mag 13(6):47–60
30. Li D, Li N, Wang J, Zhu T (2015) Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning. Knowl-Based Syst 84:214–223
31. Zhuo L, Geng Z, Zhang J, Li X (2016) Orb feature based web pornographic image recognition. Neurocomputing 173:511–517
32. Han J, Awad G, Sutherland A (2009) Automatic skin segmentation and tracking in sign language recognition. IET Comput Vis 3(1):24–35
33. Zhu Q, Wu C-T, Cheng K-T, Wu Y-L (2004) An adaptive skin model and its application to objectionable image filtering. In: Proceedings of the 12th annual ACM international conference on multimedia, pp 56–63
34. Abdullah-Al-Wadud M, Chae O (2008) Skin segmentation using color distance map and water-flow property. In: 2008 the fourth international conference on information assurance and security, pp 83–88. IEEE
35. Chen W-C, Wang M-S (2007) Region-based and content adaptive skin detection in color images. Int J Pattern Recognit Artif Intell 21(05):831–853
36. Mahmoodi MR, Sayedi SM (2014) Boosting performance of face detection by using an efficient skin segmentation algorithm. In: 2014 6th international conference on information technology and electrical engineering (ICITEE), pp 1–6. IEEE
37. Kim Y, Hwang I, Cho NI (2017) Convolutional neural networks and training strategies for skin detection. In: 2017 IEEE international conference on image processing (ICIP), pp 3919–3923. IEEE
38. Lei Y, Yuan W, Wang H, Wenhu Y, Bo W (2016) A skin segmentation algorithm based on stacked autoencoders. IEEE Trans Multimed 19(4):740–749
39. Dourado A, Guth F, Campos TE, Weigang L (2019) Domain adaptation for holistic skin detection. arXiv preprint arXiv:1903.06969
40. Ma C-H, Shih H-c (2018) Human skin segmentation using fully convolutional neural networks. In: 2018 IEEE 7th global conference on consumer electronics (GCCE), pp 168–170. IEEE
41. Roy K, Sahay RR (2021) A robust multi-scale deep learning approach for unconstrained hand detection aided by skin segmentation. The visual computer, pp 1–25

42. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134

43. Arsalan M, Kim DS, Owais M, Park KR (2020) Or-skip-net: outer residual skip network for skin segmentation in non-ideal situations. Expert Syst Appl 141:112922

44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems 30

45. Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130

46. Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C (2018) Disan: directional self-attention network for rnn/cnn-free language understanding. In: Proceedings of the AAAI conference on artificial intelligence, vol 32

47. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning, pp 7354–7363. PMLR

48. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19

49. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3146–3154

50. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision, pp 818–833. Springer

51. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: a deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147

52. Li H, Xiong P, Fan H, Sun J (2019) Dfanet: Deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9522–9531

53. Wang Y, Zhou Q, Liu J, Xiong J, Gao G, Wu X, Latecki LJ (2019) Lednet: a lightweight encoder-decoder network for real-time semantic segmentation. In: 2019 IEEE international conference on image processing (ICIP), pp 1860–1864. IEEE

54. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258

55. Feng X, Gao X, Luo L (2020) Hlnet: a unified framework for real-time segmentation and facial skin tones evaluation. Symmetry 12(11):1812

56. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5693–5703

57. Poudel RP, Liwicki S, Cipolla R (2019) Fast-scnn: fast semantic segmentation network. arXiv preprint arXiv:1902.04502

58. Furtado P (2021) Testing segmentation popular loss and variations in three multiclass medical imaging problems. J Imaging 7(2):16

59. Tan WR, Chan CS, Yogarajah P, Condell J (2011) A fusion approach for efficient human skin detection. IEEE Transactions on Industrial Informatics 8(1):138–147

60. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495

61. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440

62. Hasan MK, Dahal L, Samarakoon PN, Tushar FI, Martí R (2020) Dsnet: automatic dermoscopic skin lesion segmentation. Comput Biol Med 120:103738

63. Khoreva A, Benenson R, Hosang J, Hein M, Schiele B (2017) Simple does it: weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 876–885

64. Germi SB, Rahtu E (2022) Enhanced data-recalibration: utilizing validation data to mitigate instance-dependent noise in classification. In: Image analysis and processing–ICIAP 2022: 21st international conference, Lecce, Italy, May 23–27, 2022, proceedings, Part I, pp 621–632. Springer

65. Kovac J, Peer P, Solina F (2003) Human skin color clustering for face detection. In: The IEEE region 8 EUROCON 2003. Computer as a tool, vol 2, pp 144–1482