Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Two-Stream U-Net for Robust Skin Detection via Color and Texture Integration

Abdelkrim Sahnoune

asahnoune@usthb.dz

University of Sciences and Technology Houari Boumediene

Djamila Dahmani

University of Sciences and Technology Houari Boumediene

Saliha Aouat

University of Sciences and Technology Houari Boumediene

Abderrahmane Abdennouz

University of Sciences and Technology Houari Boumediene

Idir Timsiline

University of Sciences and Technology Houari Boumediene

Research Article

Keywords: Skin detection, Texture, Color, Two-Stream U-NET, Hierarchical Attention

Posted Date: September 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8055693/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Two-Stream U-Net for Robust Skin Detection via Color and Texture Integration

Abdelkrim Sahnoune[1*], Djamila Dahmani[1], Saliha Aouat[1], Abderrahmane Abdennouz[1], Idir Timsiline[1]

[1*]Computer Science Faculty, University of Science and Technology Houari Boumediene, Algiers, Algeria.

*Corresponding author(s). E-mail(s): asahnoune@usthb.dz;
Contributing authors: ddahmani@usthb.dz; saliha.aouat@usthb.edu.dz; aabdennouz@etu.usthb.dz; itimsiline@etu.usthb.dz;

Abstract

Skin detection plays a vital role in various computer vision applications, including face detection, gesture tracking, security surveillance, and medical diagnosis. However, traditional methods that rely solely on color information exhibit significant limitations under challenging conditions such as varying illumination and background regions with skin-like colors where chromatic information alone is insufficient and the distinctive texture of human skin becomes crucial for accurate separation. In this work, we propose a hybrid approach for skin detection, based on a two-stream U-Net. The first stream processes color information, while the second one integrates a texture-based skin probability map. The two streams are combined using a hierarchical attention mechanism to produce the final segmentation mask. Our method was evaluated in the Pratheepan, SFA, and HGR datasets, achieving satisfactory results and demonstrating its effectiveness and generalizability under various conditions. The proposed approach improves accuracy while reducing texture-related computational costs by relying on texture descriptor prediction rather than direct extraction. This design choice demonstrate the value of rethinking the integration of texture in modern deep learning approaches for efficient and accurate skin detection.

Keywords: Skin detection, Texture, Color, Two-Stream U-NET, Hierarchical Attention

1 Introduction

Human skin is undeniably an inseparable part of our identity as humankind. Its unique properties distinguish us from other species and play an essential role in our survival and social interactions. Skin is one of the elements that help us identify each other and serves as a visible indicator of our health and physical condition. As a result, skin detection occupies an important place in computer vision, with a growing need in various fields, particularly human-related ones. These include face detection [1], gesture recognition [2], people tracking [3], human-computer interaction [4], content filtering [5], and medical diagnosis [6, 7]. Skin detection defined as the identification of skin pixels or regions in an image or video has emerged as a key preprocessing step in many of these systems and in certain contexts, can even contribute directly to safety and well-being.

Numerous skin detection methods have been proposed, ranging from simple thresholding techniques to complex deep learning models. These methods differ mainly in terms of features used and computational cost. Among the various visual features used for skin detection, color has historically been the most widely employed. Its inherent simplicity and low computational cost established it as the cornerstone of most approaches. Indeed, under controlled lighting and with noncomplex backgrounds, color-based methods can achieve highly accurate results. However, color's inherent simplicity also constitutes its primary vulnerability. Color-based methods become significantly less effective under luminosity variations, complex backgrounds containing objects with skin-like colors (such as wood, sand, or certain fabrics), and skin tone variations related to ethnic diversity and individual characteristics. These limitations severely compromise their robustness in unconstrained environments.

To address these limitations, texture has been explored as a complementary [8, 9]—or in some cases, alternative—feature for skin detection. Texture captures the spatial arrangement of skin microstructures such as pores, fine wrinkles, and hair, which tend to remain stable under varying illumination. When integrated appropriately, texture exhibits greater robustness in scenarios where color-based methods struggle and provides a deeper understanding of surface-level patterns. However, adopting texture as a key feature comes at a cost, especially when using effective texture modeling techniques such as the Gray-Level Co-occurrence Matrix (GLCM) and Haralick features. Its extraction is often computationally demanding, particularly on high-resolution images. Furthermore, many texture descriptors are sensitive to scale variations, occlusions, and local rotations. Thus, while texture provides enhanced robustness, it often introduces considerable processing complexity.

To overcome the aforementioned limitations and benefit from the strengths of both color and texture, we present in this paper a hybrid approach that synergistically combines these two features. Our method is based on a two-stream U-Net architecture. The first stream operates on color by processing YCbCr channels, while the second focuses on texture information. Rather than relying on implicit texture learning within Convolutional Neural Networks (CNNs), we explicitly incorporate well-established handcrafted texture descriptors as an additional source of information. However, instead of computing them directly, which would be computationally expensive, we train a lightweight predictive model to estimate these descriptors efficiently.

The predicted texture features, generated through this novel and computationally efficient texture pathway, are then passed to a Support Vector Machine (SVM) that produces the skin probability map, which constitutes the output of the second stream. Finally, the outputs of both streams are fused using a hierarchical attention mechanism that weights multi-scale features to optimize segmentation. This design combines the efficiency of color-based processing with the robustness of texture-based analysis to produce the final segmentation. Our main contributions are:

1. We propose the integration of texture as an explicit complementary attribute to color within a two-stream U-Net architecture. The two modalities are jointly exploited via a hierarchical attention mechanism that modulates their contributions, enabling the network to balance color and texture cues during segmentation adaptively.
2. We address the high computational cost of traditional texture feature extraction by proposing a predictive texture estimation module. A lightweight model is trained to estimate multi-descriptor texture features directly from the input, providing a practical and efficient solution for incorporating texture information without the overhead of explicit computation.
3. We employ a knowledge-guided learning strategy in which we leverage a classical machine learning model, a Support Vector Machine, as an external source of knowledge to guide the deep segmentation network. By generating a texture-based skin probability map, this model enriches the deep learning framework with prior, handcrafted knowledge, effectively bridging traditional and modern learning paradigms.

The remainder of this paper is organized as follows. Section 2 reviews skin detection related works. Section 3 presents a detailed description of the proposed method. Section 4 describes the experimental setup and reports results on benchmark datasets. Section 5 concludes the paper and discusses directions for future work.

2 Related works

In this section, we provide an overview of existing skin detection methods. Over the years, a wide range of methods has been proposed to solve this problem, reflecting its importance in various computer vision tasks. Skin detection methods can be broadly classified into four key categories, each offering particular advantages and confronting specific limitations.

2.1 Color-Based Methods

Color-based skin detection methods rely on the chromatic characteristics of human skin to distinguish skin pixels from other regions of an image. These approaches are widely adopted due to their simplicity, and computational efficiency making them suitable for real-time applications. Typically, they operate by defining threshold ranges within specific color spaces where skin tones are expected to cluster. Pixels falling within these predefined ranges are classified as skin pixels.

Early skin detection methods relied on simple thresholding in various color spaces [1, 10, 11]. Although effective in controlled environments, they remain sensitive to illumination changes and ethnic diversity. Recent studies have explored strategies such as adaptive thresholds and color normalization to overcome these challenges.

Brancati et al. [12] proposed a dynamic rule-based skin detection method in the YCbCr color space designed to adapt to varying illumination conditions. Their approach identifies image-specific skin clusters in the YCb and YCr subspaces and applies correlation-based rules that exploit the inverse relationship between chrominance components to classify pixels as skin or non-skin.

Samson and Lu [13] introduced a fast spatial model combining improved color thresholding in RGB and HSL spaces, optimized for dark skin, with The Packed k-dimensional Tree (PKT). This hierarchical structure partitions the image into subregions and performs interpolation on only a few strategically selected pixels per cell for classification.

Satheeskumar [14] proposed a face detection method based on skin segmentation in the L*a*b* color space, exploiting the chromatic components a and b, independently of the luminance component (L*). Skin regions are extracted using global thresholding, effectively reducing the impact of lighting variations. Final face localization is achieved through geometric rules. Experimental results suggest that the method performs reliably under varying illumination conditions.

2.2 Texture-Based Methods

Texture-based skin detection methods exploit the spatial patterns specific to skin regions to distinguish them from other parts of the image. Unlike color-only approaches, these techniques analyze surface characteristics such as granularity, roughness, or local regularity, making them particularly effective in differentiating skin from backgrounds with similar chromatic properties.

Wang et al. [15] integrated texture features such as contrast and homogeneity, derived from the GLCM with with YCbCr and RGB color models. This combination improved detection rates and enhanced the separation of skin and non-skin regions, particularly in images with complex backgrounds.

Sarkar et al [16] proposed a novel skin detection approach that leverages prior geometric knowledge from face detection to localize initial skin regions. The framework then combines texture-based features including Local Binary Patterns (LBP) and lacunarity with grayscale intensity statistics to identify new candidate skin patches. A superpixel-based region-growing technique further refines segmentation by propagating labels to adjacent regions.

2.3 Machine Learning-Based Methods

Machine learning-based methods use statistical models trained on annotated data to automatically distinguish skin regions in an image. Unlike fixed-rule approaches, these techniques learn decision boundaries from examples often integrating multiple features such as color, texture, and local shape.

Khan et al. [17] proposed a Random Forest–based skin detection method in the IHLS color space. They evaluated their model against six other classifiers including Bayesian network, SVMs, neural networks, and decision trees on a dataset of 8,991 images with pixel-level ground truth, achieving higher accuracy.

Bush et al. [18] proposed an adaptive algorithm based on Adaptive Neuro-Fuzzy Inference Systems (ANFIS), combining the strengths of neural networks and fuzzy logic. Their method operates directly in the RGB color space, avoiding any color space conversion. It integrates fuzzy c-means (FCM) clustering with Sugeno-type inference to automatically generate fuzzy rules and membership functions for skin classification.

Chakraborty and Bhuyan [19] proposed an unspurvised adaptive discriminative analysis method for skin segmentation. By extracting color, texture, and spatial cues and applying adaptive region growing, they derive image-specific discriminative feature sets, which are then used for pixel-wise skin classification.

2.4 Deep Learning-Based Methods

Deep learning-based methods have recently revolutionized skin detection by enabling discriminative features to be learned directly from raw images. Unlike conventional approaches that rely on manually selected features, deep networks automatically extract relevant patterns through multiple layers of abstraction, effectively capturing color, texture, shape, and spatial context. As a result, they have become the dominant paradigm in recent skin detection research.

Tarasiewicz et al. [20] proposed Skinny, a lightweight and efficient U-Net–based architecture for skin detection. To address the limited spatial context of the original U-Net [21], they introduced an optimized design incorporating inception blocks and dense connections, enabling broader multiscale analysis while reducing the number of parameters. In addition to being computationally efficient, Skinny demonstrates greater robustness to variations in shade and illumination, and notably outperforms the classic U-Net and manual feature-based methods.

Ben Salah et al. [22] proposed a two-step method for skin detection utilizing a patch-based CNN. In their approach, the first step involves training a CNN to classify individual image patches as either skin or non-skin. Subsequently, this trained model is integrated into a scanning algorithm that processes the entire image to generate the final skin region mask.

Leite et al. [23] proposed a hybrid approach that combines Mask R-CNN, BodyPix, and U-Net to segment skin pixels in images captured under uncontrolled conditions. The method first uses Mask R-CNN to detect regions of interest (i.e., people), followed by BodyPix to segment body parts such as the face, arms, torso, and legs. Finally, a U-Net trained on a heterogeneous dataset classifies skin and non-skin pixels by merging HSV and CIELab color spaces.

Hashemifard et al. [24] proposed a weakly supervised segmentation method that integrates two guided attention modules, namely Body Attention and Skin Attention. These modules use contextual information to guide the network toward relevant body and skin regions.

Khosravi et al. [25] employed the Chimp Optimization Algorithm (ChOA), inspired by the social behavior of chimpanzees, to optimize the initial weights and biases of a

convolutional neural network (CNN) prior to training. This hybrid approach combining ChOA with gradient descent helps the network avoid convergence to local minima and improves training efficiency. The resulting model achieves a very high correct detection rate and outperforms several recent skin detection methods.

Kuban and Kawulok [26] proposed a hierarchical assembly of CNNs for skin segmentation. Their approach combines four CNNs: one trained on RGB images to capture color-based features, one trained on grayscale images to extract textural features, and two others guided by regions pre-segmented using a Bayesian classifier. All four specialized models use the Skinny network as backbone. Each model generates a skin probability map, which are subsequently merged by a second CNN to produce the final segmentation.

2.5 Skin detection methods: Challenges and contributions

Although color-based methods are simple to implement and computationally efficient, they suffer from several limitations. They are particularly sensitive to lighting conditions and image quality, as well as to the presence of objects with skin-like tones. Moreover, the wide diversity of skin pigmentation across different ethnicities makes it difficult to define universal thresholds that are perform reliably in all situations.

Texture-based approaches are often used to complement color-based methods, particularly in hybrid systems, as texture alone may not provide sufficient discrimination especially at low resolutions. However, in high-quality images or situations where color becomes unreliable due to difficult lighting or occlusions, texture information can play a crucial role. The main drawback remains their high computational cost, which limits their applicability in real-time scenarios.

Machine learning–based methods represent a notable advancement over traditional rule-based approaches. By learning decision boundaries from annotated data, they can integrate multiple features to enhance skin detection performance. These methods are generally more adaptable and interpretable than thresholding techniques, while requiring less data and computational power than deep learning models. However, their generalization ability remains limited under highly variable conditions, as they often rely on manually engineered features and lack the hierarchical representation learning of deep neural networks.

Deep learning-based methods significantly outperform the previously mentioned approaches, particularly under real-world conditions such as lighting variations, occlusions, ethnic diversity, and complex backgrounds. They also exhibit greater robustness to image imperfections including compression artifacts, noise, and blur. However, these methods come with notable limitations: they require large volumes of annotated data to generalize effectively. They demand substantial computational resources (often GPU-based). In addition, their decision processes are typically difficult to interpret.

In this context, our method aims to address key challenges in skin detection, including sensitivity to lighting variations and image quality, confusion with objects of similar color, and most importantly, the high computational cost associated with texture feature extraction. To this end, we propose a two-stream U-Net that combines color information with learned texture descriptors through a hierarchical attention mechanism.

3 Proposed method

In this section, we provide a detailed description of our proposed skin detection approach. Our approach adopts a two-stream architecture based on U-Net, with each stream dedicated to processing a specific type of information: one stream focuses on chromatic features derived from the original image, while the other handles textural information represented by a skin probability map.

Injecting textural features directly into the U-Net bottleneck would inevitably result in a loss of spatial information and fine-grained structure present in the original input. Additionally, merging texture and color information at this stage may cause the color features to dominate, limiting the effective contribution of texture. For this reason, we adopt a dedicated texture stream, constructed from a reduced set of selected descriptors, allowing more effective modulation of its contribution through attention-based fusion.

This design enables the model to preserve spatial resolution while leveraging the discriminative power of texture-based features.

As illustrated in Fig. 1, the proposed architecture consists of the following main components:

- A first encoder stream that processes the YCbCr representation of the input image.
- A second encoder stream that processes a skin probability map generated by an SVM classifier trained on predicted texture descriptors.
- A set of attention gates that fuse information from both streams.

The outputs from the two encoders are integrated through a unified decoder to generate the final segmentation. This architecture enables the progressive fusion of information at various resolution levels, thereby ensuring optimal preservation of fine details and accurate skin contours.

In the remainder of this section, we describe the components of our method in the following order: the color processing stream, the texture-based probability map stream, the attention-based fusion module, and the final segmentation output.

3.1 Color stream

To enhance chromatic representation and improve model performance, we first apply a preprocessing pipeline to the original RGB image. This step aims to reduce the impact of illumination variability while preserving meaningful color information.

The input image is first converted to the YCbCr color space, which separates luminance (Y) from chrominance (Cb and Cr). This separation increases robustness to lighting changes, making YCbCr a more suitable representation for skin detection than RGB.

Next, we apply a slight Gaussian blur to attenuate local noise without altering important structures. To further enhance contrast, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) specifically on the Y channel. This step improves local contrast while preventing over-amplification of noise, allowing better preservation of subtle textural cues in skin regions.

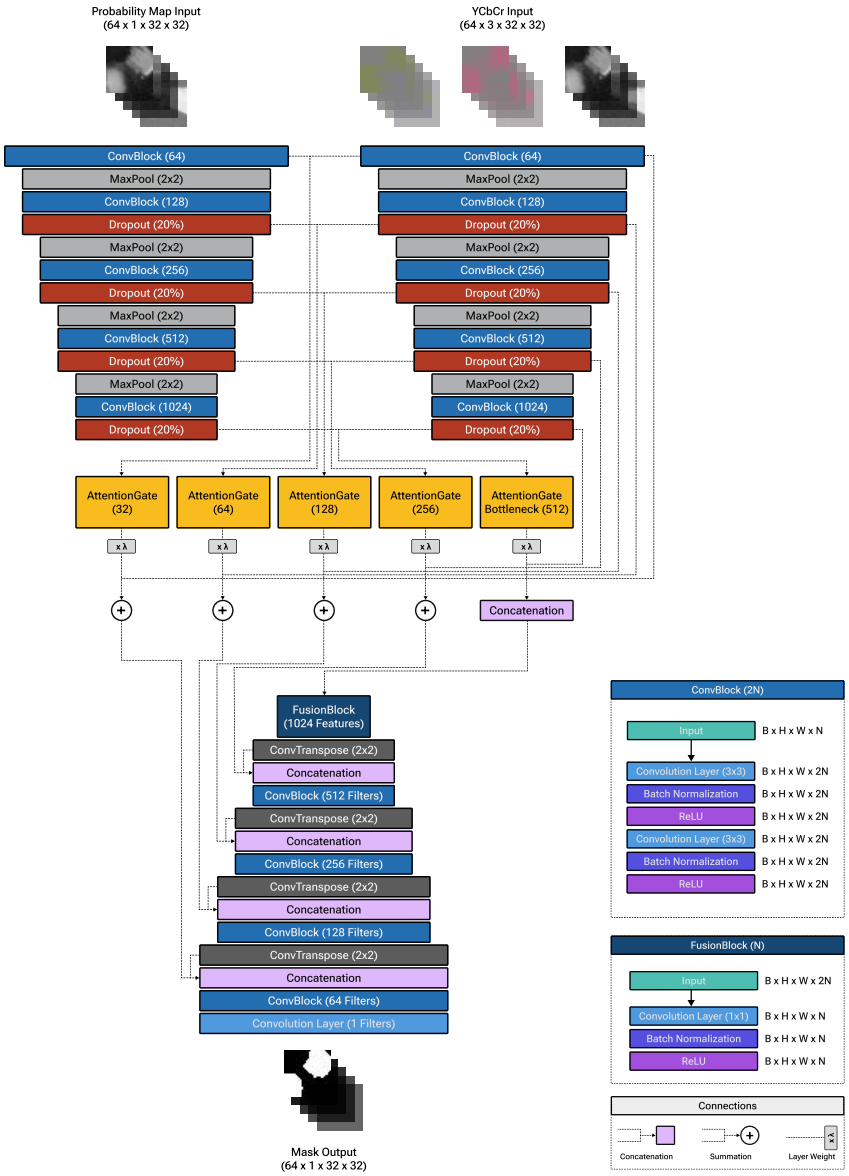


Fig. 1 Overall system architecture of our Two Stream UNet with hierarchical attention mechanism

By applying this preprocessing process, we preserve essential information while normalizing lighting variations that may disrupt segmentation. The resulting chromatic representation, enhanced by adaptive contrast normalization, encodes color-based skin cues and serves as the input to the first encoder stream of our U-Net architecture.

3.2 Texture stream

The texture stream of our architecture comprises two main components: (1) the prediction of texture descriptors using a lightweight CNN, and (2) the generation of a skin probability map using an SVM classifier trained on these descriptors.

3.2.1 Texture Descriptors Prediction

As a first step, we examine the structural and visual characteristics of human skin to identify consistent patterns that persist across different skin tones and help distinguish skin from visually similar materials. Skin texture is inherently complex and cannot be fully captured using a single method. It is a multifaceted visual property that varies across scale and is influenced by local patterns, intensity contrast, frequency, and orientation. However, each descriptor tends to focus on a specific aspect while neglecting others. Therefore, we propose combining statistical, structural, and transform-based methods, as they offer complementary descriptors that together enable a richer and more comprehensive representation of skin texture. To this end, we consider the following descriptors.

- Haralick descriptors: Derived from the Gray-Level Co-occurrence Matrix (GLCM), these second-order statistics capture spatial relationships between pixel intensities. These descriptors are effective in characterizing subtle textural differences between skin and non-skin regions, by highlighting contrast, homogeneity, and other structural differences. The characteristics considered are the following. Variance, Entropy, Energy, Homogeneity, Third-Order Moment, Inverse Variance, and Cluster Shade.
- Local Binary Patterns (LBP): LBP encodes local texture by thresholding each pixel's neighborhood, effectively capturing uniform patterns typical of skin regions. Its robustness to changes in illumination improves its reliability in various lighting conditions [27].
- Gabor filters: These filters extract localized frequency and orientation information, allowing modeling of repetitive patterns and directional textures common in skin microstructure, such as pores and wrinkles.
- In addition, a close examination of the skin tissue reveals that the interface between the dermis and epidermis exhibits a fractal-like structure. This structural folding, common in biological systems, increases surface area and improves functional adhesion. To quantify this, we compute the Fractal Dimension (FD), which measures the self-similarity and roughness of a surface. We calculate the Fractal Dimension (FD) using its relationship with the Hurst exponent H, derived from the fractional Brownian motion. The Hurst exponent is estimated using wavelet transforms as presented in [28], as follows:

$$H \approx \frac{\log |X * \psi_{j_1}(t)|}{\log 2} \tag{1}$$

where $X * \psi_{j_1}(t)$ denotes the wavelet transform of the signal X at scale j_1. Once H is estimated, the Fractal Dimension is computed as:

$$FD = n + 1 - H \tag{2}$$

with $n = 2$ for 2D image patches, yielding:

$$FD = 3 - H \tag{3}$$

This formulation enables us to capture the skin's space-filling complexity using a single scalar value.

Therefore, to generate a texture descriptor set, we divide the image into 8×8 patches and compute Haralick, LBP, Gabor, and Fractal Dimension features for each patch. However, this extraction process is computationally expensive, particularly when applied to large datasets or during inference. To overcome this limitation, we adopt a predictive strategy. Instead of computing texture descriptors explicitly during inference, we train a lightweight CNN to predict the texture descriptors from raw image patches. For training, we manually extract a set of 25 texture features per 8×8 patch, comprising:

- 7 Haralick features computed from GLCM with distance $d = 1$ and angles $\theta = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$.
- 10 LBP features.
- 4 Gabor-based features with orientations uniformly distributed from 0 to π (8 orientations).
- 4 Hurst exponents using Daubechies wavelet families (db1, db4, db8, and HAAR).

The Hurst exponent is used directly, as it offers equivalent information to the fractal dimension (FD is simlpy a linear transformation of the Hurst exponent $FD = 3 - H$) while avoiding redundant computation.

All extracted descriptors are stored as structured input data to serve as the ground truth for CNN training. Our predictive CNN illustrated in Fig. 2 processes 8×8 patches, which represents a good compromise between preserving spatial details and computational efficiency. Its architecture consists of:

- Three convolutional layers with no padding and small kernel sizes for gradual spatial reduction and feature abstraction.
- Global adaptive average pooling to aggregate patch information into a compact 64-dimensional representation.
- Two fully connected layers with dropout (0.4, 0.2) for regularization and final regression of the 25 texture features.

This design ensures controlled spatial compression while maintaining enough representational power to capture relevant patterns. Batch normalization and dropout help improve generalization and training stability. Once trained, the CNN can estimate texture descriptors for new patches without explicit feature computation, thus drastically reducing inference time.

To validate the CNN predictor, we evaluated its predictions against the traditionally computed texture descriptors used during training. The resulting regression

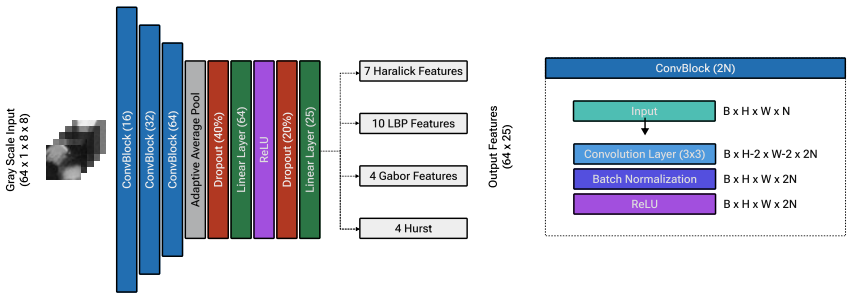


Fig. 2 Predictive CNN architecture for estimating 25 texture descriptors from 8×8 skin image patches

errors remained consistently low across all feature types, indicating that the CNN successfully learned to approximate the original handcrafted descriptors.

3.2.2 SVM Probability Map Generation

The texture descriptors predicted by the CNN are passed to a Support Vector Machine (SVM) classifier to generate a skin probability map, representing the likelihood that each patch belongs to the skin class. This map encodes texture-based information in a spatial format suitable for integration into the U-Net architecture. To handle the non-linear nature of the feature space, we use an SVM with a Radial Basis Function (RBF) kernel, approximated using Random Fourier Features. The approximation projects the original feature vector x into a randomized feature space $\phi(x)$, such that the kernel function is estimated as:

$$K(x, x') \approx \phi(x)^T \phi(x'), \tag{4}$$

where $K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$ is the original RBF kernel. The parameter γ controls the locality of influence of support vectors, and thus the decision boundary's flexibility. To efficiently handle large images, we adopt a three-step approach:

1. We begin by normalizing the input texture features to ensure uniform scaling for the SVM.
2. We approximate the non-linear RBF kernel with random Fourier Features, which transform the input space into a fixed dimension to reduce computation for large images.
3. We use an optimized linear SVM trained using stochastic gradient descent (SGD). This approach enables efficient learning on the transformed data obtained from the RBF kernel approximation, while automatically handling class imbalance.

The main hyperparameters are the kernel width γ and the regularization coefficient α, which controls the trade-off between model complexity and fitting. These hyperparameters are optimized through cross-validation along with the number of random components used in the kernel approximation.

Once trained, the SVM produces a texture-based skin probability map indicating for each pixel the likelihood of belonging to the skin class. To reduce block artifacts inherent to patch-wise processing, we apply bilinear interpolation for spatial smoothing. The resulting map serves as input to the second encoder stream of our U-Net

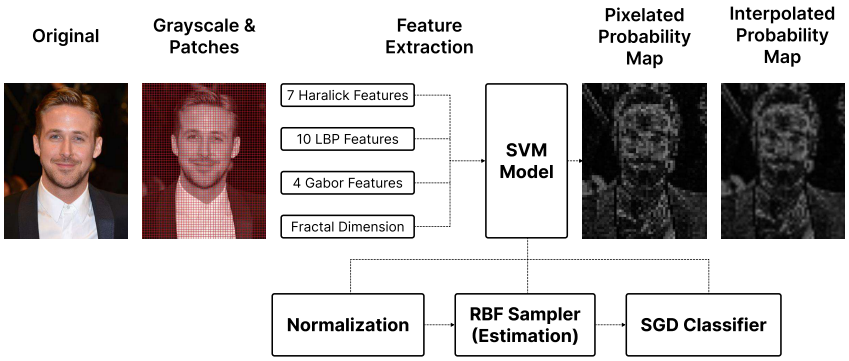


Fig. 3 Example of probability map generated by our SVM model

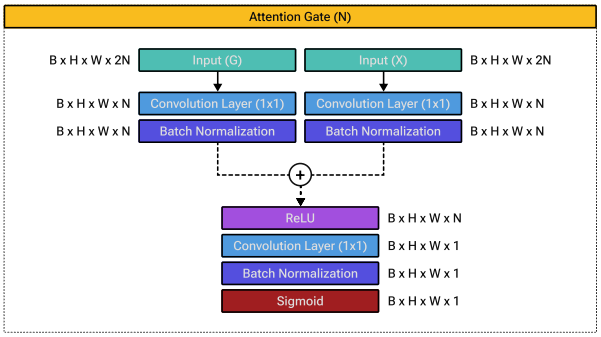


Fig. 4 Architecture of the hierarchical attention gate

architecture, complementing chromatic information in guiding the final segmentation. Fig. 3 illustrates an example of a probability map generated by our SVM model.

3.3 Hierarchical Attention Fusion

Once the chromatic and texture-based inputs are fed into the network, they are processed independently through two parallel encoders. The subsequent step is to determine an effective way to merge the two streams. This requires adjusting their respective contributions so that they complement each other, thereby improving U-Net's output. To this end, we use a hierarchical attention mechanism that modulates the contribution of each stream: the color path, which captures appearance-related features, and the texture path, which focuses on structural patterns.

Each attention gate (see Fig. 4) operates in three steps:

1. Independent transformation of features through 1×1 convolutions and batch normalization.
2. Element-wise addition and generation of an attention map through a specialized convolution.
3. Sigmoid activation to produce modulation coefficients ranging in $[0, 1]$ that scale the contribution of the texture stream.

This mechanism can be formulated as follows:

$$\mathcal{A}(f_c, f_t) = \sigma\left(\psi\left(\text{ReLU}\left(\phi_c(f_c) + \phi_t(f_t)\right)\right)\right) \odot f_t \tag{5}$$

$$f_{\text{fused}} = f_c + \lambda \cdot \mathcal{A}(f_c, f_t) \tag{6}$$

where f_c and f_t denote the color and texture feature maps, ϕ_c and ϕ_t are 1×1 convolutions with batch normalization, ψ is a convolution that reduces activation to a single channel attention map, σ is the sigmoid function, and $\lambda \in [0, 1]$ is a fixed parameter controlling the global influence of texture. The operator $\odot$ denotes element-wise multiplication.

The mechanism adopts a multilevel approach that operates simultaneously at several U-Net resolutions. At the level of residual connections (64, 128, 256, and 512 channels), the gates ensure progressive fusion that preserves fine details. At the bottleneck (1024 channels), a more global fusion combines high-level semantic representations.

A global parameter λ controls the overall influence of the texture stream, allowing the network to adapt to different types of images and acquisition conditions. This hierarchical attention mechanism plays a critical role in situations where chromatic information alone is insufficient, such as the presence of backgrounds or objects with skin-like colors. By dynamically relying on texture cues, the model can correct and refine its predictions. The enriched features are then processed by the U-Net decoder to generate the skin probability map, which is thresholded at $\theta = 0.5$ to produce the final binary segmentation mask.

4 Experimental results

In this section, we present the experiments conducted to evaluate our proposed skin detection method, which integrates color and texture information within a deep learning framework. We describe the experimental setup and provide a quantitative comparative evaluation, as well as qualitative results. These results offer insights into the performance and efficiency of the proposed approach.

4.1 Datasets

For evaluation, we selected publicly available datasets that are representative, well annotated, and widely used in the literature, enabling meaningful comparison with existing approaches. These datasets encompass a wide range of scenarios for skin pixel detection, including variations in ethnicity, illumination, background complexity, and image quality. **Hand Gesture Recognition (HGR) dataset :**[29–31] Comprising three subsets (HGR1, HGR2a, and HGR2b) with a total of 1,558 images and corresponding binary masks, this dataset offers controlled samples of hand skin in various positions and gestures. It is particularly useful for evaluating performance on hand-specific skin detection tasks.

Pratheepan dataset :[32] Designed specifically for skin segmentation research, this dataset contains 78 images with manually annotated ground truth masks. It

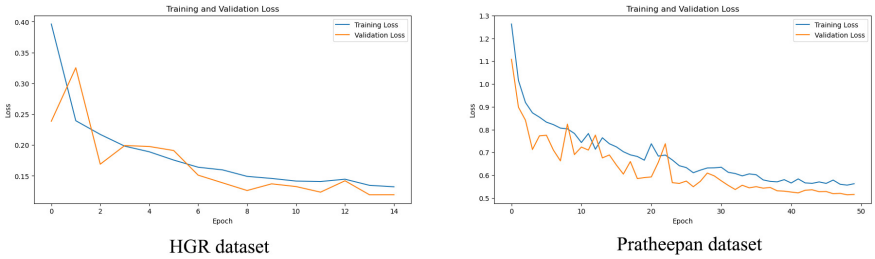


Fig. 5 Training and validation loss evolution for U-Net model on HGR (left) and Pratheepan (right) datasets.

includes challenging conditions such as complex backgrounds, shadows, and lighting variations, making it particularly suitable for assessing robustness in real-world settings.

Skin Face Annotations (SFA) dataset :[33] This dataset contains 1,118 high-quality face images with pixel-accurate annotations. It covers diverse facial expressions, poses, and demographic profiles under different lighting conditions. Its emphasis on fine delineation of facial skin regions, enabling rigorous evaluation of facial skin segmentation accuracy.

4.2 Implementation details

Our approach was implemented using PyTorch, Scikit-learn, and OpenCV, and experiments were conducted on a NVIDIA GeForce GTX 1050 Ti GPU (4 GB GDDR5). **CNN-based Texture Feature Prediction:** The CNN was trained using the Mean Squared Error (MSE) loss and the Adam optimizer with a learning rate of 0.003, weight decay of 0.00005, and β parameters set to (0.9, 0.999). A ReduceLROnPlateau scheduler (mode = 'min', patience = 3, factor = 0.5, minimum learning rate =1×10^{-5}) was used. Training was performed for 50 epochs on all datasets. **SVM-based Skin Probability Map Generation:** The SVM used hinge loss with L2 regularization and balanced class weights to address class imbalance. The maximum number of iterations was set to 1000 with a tolerance of 1×10^{-4}. For the RBF kernel approximation, we employed random Fourier features with parameters $\alpha = 1 \times 10^{-6}$, $\alpha = 0.01$, and 300 components. For Pratheepan, the number of components was increased to 2000 to improve generalization on the limited training set. **U-Net Training:** The U-Net was trained using a composite loss function combining Binary Cross-Entropy with Logits Loss and Dice Loss, in order to jointly optimize pixel-wise accuracy and segmentation quality. The Adam optimizer was configured with a learning rate of 5×10^{-4} and weight decay of 1×10^{-5}, with the same ReduceLROnPlateau scheduler as the CNN. Training was conducted for 15 epochs for SFA and HGR datasets, and 50 epochs for Pratheepan dataset. Fig. 5 displays the training and validation loss curves for the HGR and Pratheepan datasets, illustrating stable convergence with no signs of overfitting, which validates the model's ability to generalize well.

4.3 Evaluation metrics

To evaluate the performance of our approach, we rely on classical metrics derived from the confusion matrix obtained by comparing the predicted segmentation masks against the ground truth. These metrics are widely adopted in skin detection research, enabling direct and fair comparisons with existing methods. The metrics are defined as follows:

Precision, measures the proportion of pixels predicted as skin that are actually skin pixels.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

Recall, measures the model's ability to identify skin pixels among all the skin pixels present.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

Specificity, measures the model's ability to correctly recognize non-skin pixels.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

F1-Score, represents the harmonic mean between Precision and Recall, reflecting a balance between these two measures.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Detection Pixel Rate of Skin (DPRS), represents a global performance indicator that combines Precision, Recall, and Specificity.

$$\text{DPRS} = \sqrt{(1 - \text{Precision})^2 + (1 - \text{Recall})^2 + (1 - \text{Specificity})^2} \tag{11}$$

Where TP is the number of pixels correctly classified as skin, TN the number of pixels correctly classified as non-skin, FP the number of pixels incorrectly classified as skin, and as FN the number of pixels incorrectly classified as non-skin.

4.4 Ablation studies

In order to study the contribution of each component in our skin detection method, we conduct an ablation study centered on the following aspects: 1) the texture stream, 2) the SVM-based probability mapping, 3) the CNN-based descriptor prediction. The experiments were conducted on the Pratheepan dataset, where each configuration focused on isolating a specific factor to assess its impact on skin detection performance. The results obtained are detailed and analyzed in the subsequent subsections.

4.4.1 Effect of Combining Texture and Color Features

To evaluate the contribution of texture information to overall performance, we compare the values of the selected metrics obtained by training the architecture with

Table 1 Performance comparison of color, texture, and combined features in skin detection

Configuration	Precision	Recall	F1-Score	Specificity	DPRS
Color only	0.5639	0.9151	0.6978	0.8525	0.4681
Texture only	0.7055	0.9212	0.7990	0.9198	0.3153
Color + Texture	**0.7796**	**0.9381**	**0.8515**	**0.9228**	**0.2415**

Table 2 Performance comparison of texture integration strategies: raw descriptors, SVM mapping, and attention-based fusion

Configuration	Precision	Recall	F1-Score	Specificity	DPRS
Texture U-Net (raw descriptors)	0.7796	0.9381	0.8515	0.9228	0.2415
Texture U-Net (SVM probability map)	0.7955	**0.9484**	0.8652	0.9257	0.2236
Two-Stream U-Net (SVM + attention)	**0.8607**	0.9472	**0.9019**	**0.9620**	**0.1537**

three configurations: (1) color features only, (2) texture features only, and (3) the combination of both. Table 1 reports the results for these configurations.

Table 1 shows that texture features alone provide substantially better performance than color features across all metrics. The precision gain exceeds 14% (0.7055 vs. 0.5639), highlighting the discriminative power of texture to distinguish skin from challenging backgrounds. Combining color and texture further improves performance, achieving the highest recall 0.9381, and specificity 0.9228. Precision also significantly increases to 0.7796, while the DPRS drops to its lowest value 0.2415, indicating reduced overall misclassification. These improved results confirm that texture features offer complementary information to color features, contributing to a more comprehensive and discriminative skin representation. Texture captures structural patterns that increase robustness to background confusion and illumination changes. Simultaneously, color information captures the chromatic consistency of skin, which reinforces the detection of true skin regions and leads to a better trade-off between recall and precision.

4.4.2 Effect of SVM Probability Maps and Attention-Based Fusion

To examine the impact of texture integration strategies alongside color information, we evaluate three settings: 1) a single-stream U-Net where texture is introduced through raw descriptors, 2) a single-stream U-Net where texture is represented by an SVM-generated probability map, and 3) our proposed two-stream U-Net, which combines color and texture via hierarchical attention. The corresponding results expressed in terms of the selected evaluation metrics are reported in Table 2.

Table 2 shows that incorporating raw texture descriptors alongside color in a single-stream U-Net achieves high recall (0.9381) but relatively low precision (0.7796). This pattern indicates the model retrieves most true skin pixels while also producing a notable number of false positives, as reflected by the higher DPRS (0.2415). Replacing raw descriptors with an SVM-generated probability map improves precision (0.7955)

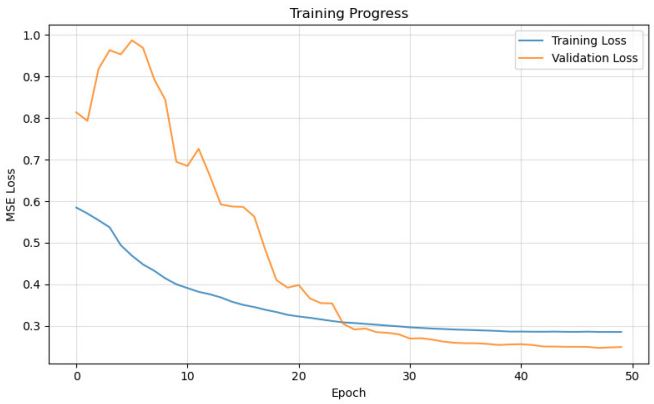


Fig. 6 Training behavior of the CNN predictor for texture descriptors (MSE)

and F1-score (0.8652), indicating that the probability map guides the network to concentrate on texture regions more likely to correspond to skin, thereby reducing noise from irrelevant patterns. The two-stream U-Net with hierarchical attention further enhances performance, achieving the highest precision (0.867), F1-score (0.9019), and specificity (0.9620), while also considerably reducing the DPRS to (0.1537). These results highlight the benefit of the two-stream architecture: processing color and texture in separate pathways preserves the integrity of each feature set. Subsequently, the attention-based fusion mechanism adaptively balances their respective contributions to maximize overall segmentation quality.

4.4.3 Effect of CNN-based Descriptor on Performance and Efficiency

To provide a comprehensive evaluation of replacing handcrafted descriptors with CNN-predicted ones, we analyze both the training behavior of the CNN predictor and, more importantly, the computational efficiency of the resulting pipeline. First, we report the training loss (MSE) to demonstrate the CNN's ability to approximate handcrafted texture descriptors. Second, to quantify computational benefits, we present inference-time measurements for both pipelines. All timings are averaged over the test set and measured on the hardware described in Subsection 4.2.

The training curves presented in Fig. 6 illustrates the evolution of the training and validation losses (MSE) across epochs. Both curves decrease steadily before stabilizing, with the validation loss converging slightly below the training loss. The best values achieved were 0.2627 for training and 0.2256 for validation, confirming that the CNN successfully learns to approximate handcrafted descriptors while maintaining good generalization.

Table 3 reports the computation time of the texture-descriptor extraction stage on the Pratheepan dataset. This dataset was chosen because it is the smallest, making it manageable with the available computational resources. All downstream stages (SVM mapping and U-Net inference) were identical across configurations and are therefore

Table 3 Descriptor-stage computation time per image (Pratheepan, $N = 78$)

Configuration	Mean (s/img)	Std (s)	Min (s)	Max (s)
Handcrafted descriptors	261.03	148.16	52.47	731.66
CNN-predicted descriptors	5.35	3.25	1.17	15.95

excluded from the timing comparison. Both setups were executed under the same machine configuration described in Subsection 4.2, ensuring a fair comparison.

As shown in Table 3, replacing handcrafted descriptors with CNN-predicted ones dramatically reduces the average computation time from 261.03 ± 148.16 seconds to only 5.35 ± 3.25 seconds. This corresponds to a $\sim 48.8\times$ speedup and a total time reduction of 97.9%. The large variance in the handcrafted configuration ($std = 148.2s, max = 731.7s$) reflects its high cost as well as its sensitivity to image size and patch count. By contrast, the CNN-based configuration shows both lower latency and tighter variance ($std = 3.25s$), demonstrating not only efficiency but also stability across samples. The large variance observed in the handcrafted configuration ($std = 148.2s$, $max = 731.7s$) reflects both its computational cost and its sensitivity to image size and patch count. In contrast, the CNN-based configuration achieves lower latency with a much tighter variance ($std = 3.25s$), demonstrating efficiency and stability across samples. These results highlight the considerable computational advantage of the CNN-based strategy, making the pipeline more suitable for time-constrained scenarios.

4.5 Comparative evaluation

To validate our approach, we compare it against established skin detection methods on the Pratheepan, HGR, and SFA datasets. Performance was assessed based on the metrics detailed in Subsection 4.3, ensuring alignment with standard practices in the field. For a fair comparison, we limited our analysis to methods that used the same datasets, specifically those referenced in [12, 20, 22, 24]. The results of this comparison are presented in Table 4, the best value per dataset and metric is highlighted in bold.

From Table 4, we can confirm that the proposed method consistently produces strong performance across all three datasets, and overall outperforms the methods it has been compared to. On the HGR dataset, the proposed method achieves the highest recall with a gain of +0.0628 and the highest F1-score with a gain of +0.0163. Our method achieves the second-best precision with a difference of −0.0307 but outperforms the method with the best precision [20] in terms of recall and F1-score. Our method also achieves the best specificity and Dprs with a gain of +0.0575 and −0.2072, respectively, compared to the method that used these metrics [12]. On the Pratheepan dataset, the proposed method attains the highest recall with a margin of +0.0377, as well as the best specificity (gain of 0.139) and Dprs (drop of 0.3612), compared to the method that used these metrics [12]. Our method achieves an F1-score on par with [24] (a gain of +0.0011). Although the work from [24] reports slightly higher precision, the overall balance across metrics favors our approach. On the SFA dataset, our proposed method achieves the highest recall while maintaining a highly competitive

Table 4 Comparative performance of skin-detection methods on the Pratheepan, HGR and SFA datasets

Dataset	Method	**Precision**	**Recall**	**F1-Score**	**Specificity**	**DPRS**
HGR	[12]	0.8938	0.7664	0.8252	0.9274	0.2667
	[20]	**0.9740**	0.9260	0.9494	/	/
	Proposed method	0.9436	**0.9888**	**0.9657**	**0.9849**	**0.0595**
SFA	[12]	0.8479	0.4206	0.5623	**0.9688**	0.5998
	[22]	**0.9400**	0.9600	**0.9500**	/	/
	Proposed method	0.9044	**0.9738**	0.9378	0.9580	**0.1076**
Pratheepan	[12]	0.5513	0.8199	0.6592	0.8230	0.5149
	[24]	**0.8924**	0.9095	0.9008	/	/
	Proposed method	0.8607	**0.9472**	**0.9019**	**0.9620**	**0.1537**

precision and F1-score. Although the method presented in [22] registers a marginally higher F1-score, our method trades the slight decrease in F1-score (-0.0122) for a recall gain of $+0.0138$. Furthermore, while a marginally higher specificity is reported in [12], our method significantly outperforms this approach in other metrics, with a drop in DPRS of 0.4922, for instance. Overall, the proposed method consistently demonstrates superiority in recall across all three datasets while maintaining competitive values in precision and F1-score. This consistent performance across diverse and challenging datasets indicates that our method prioritizes minimizing false negatives (missed skin pixels) without a significant sacrifice in precision, confirming its effectiveness for applications where complete skin pixel detection is crucial.

4.6 Qualitative results

In this subsection, we present the qualitative results of our proposed method to visually assess its performance across all three datasets and demonstrate its effectiveness on challenging scenarios such as varying lighting conditions, complex backgrounds, and diverse skin tones. These visual results complement the quantitative analysis by illustrating the method's ability to delineate skin boundaries and handle complex cases accurately. To highlight the improvement achieved by our method, we also include results from [12] method for comparison. As a rule-based approach relying solely on color information, the [12] method serves as an effective baseline to illustrate the challenges of color-only methods and demonstrate the added value of incorporating texture cues in our dual-stream framework.

Figures 7–9 show the results of our skin detection method on examples from the HGR, Pratheepan, and SFA datasets, respectively. For each figure, for each row, we show the input image, its corresponding ground-truth mask, the result from [12] method, and our predicted segmentation. Note that both ground-truth masks and predictions are binary images where white pixels represent skin regions and black pixels represent non-skin regions.

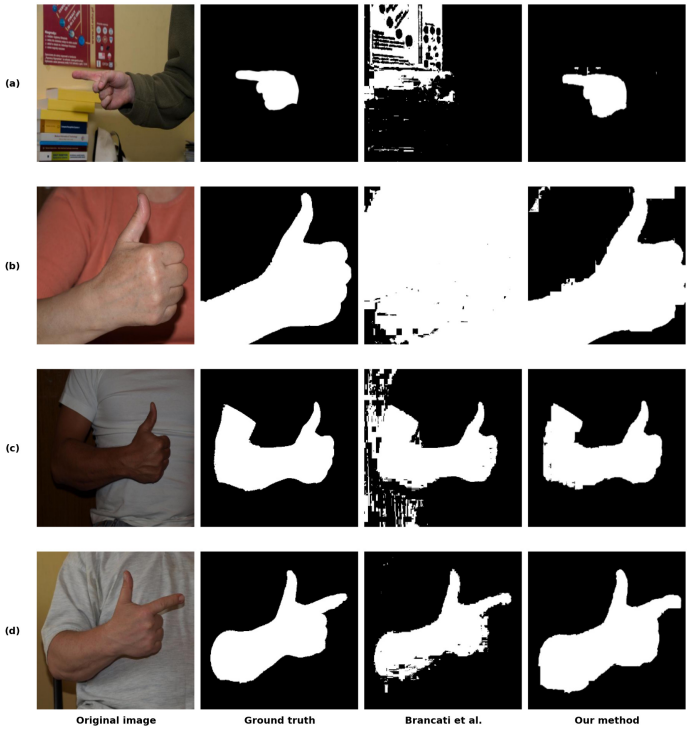


Fig. 7 Qualitative results on the HGR dataset. From left to right: input image, ground truth, [12], proposed method

From Figures 7–9, we observe that our skin detection method demonstrates remarkably robust and reliable performance overall, maintaining effectiveness under variations in lighting, backgrounds, skin tones, and at different scales. In particular, our method considerably reduces false positives by more accurately eliminating background regions incorrectly classified as skin. This trend is particularly evident in Fig. 7(b), where the approach proposed in [12] incorrectly classifies almost all the input images as skin due to clothing color similarity with skin, incorporating a large part of the background. Additionally, we observe a significant improvement in the detection of missing skin pixels, which is particularly visible in Fig. 9(c), where the method presented in [12] produces fragmented masks with large missing skin areas. The incorporation of texture information through attention-based fusion is especially beneficial in these cases, for distinguishing complex backgrounds (clothing with skin-like tones) from actual skin, thus avoiding misclassification while simultaneously preserving fine-grained details such as hands and facial contours. These results align with the quantitative performance, confirming that the proposed integration of CNN-predicted texture descriptors and color information not only improves detection rates but also produces visually more accurate and reliable segmentations. Another notable advantage of our method is its stability across different image types. Whether single-subject



Fig. 8 Qualitative results on the Pratheepan dataset. From left to right: input image, ground truth, [12], proposed method

close-ups (Fig. 9) or multi-person scenes (Fig. 8(d)), the segmentation remains coherent and contiguous, in contrast to the fragmented masks seen in [12] results. It is also worth noting that the HGR dataset contains only hands, while the SFA dataset focuses exclusively on faces, providing different skin regions with distinct textural and color characteristics. The proposed method performs reliably in both cases, confirming our method's ability to generalize across different anatomical regions.

4.7 Summary and Discussion

The thorough evaluation confirms the effectiveness of our proposed method for skin detection. Across all three datasets, quantitative results consistently demonstrate superior recall performance, accompanied by very competitive precision and specificity. Visual analysis further demonstrates the method's robustness under diverse and challenging conditions. Our method's consistent prioritization of minimizing false negatives significantly contributes to the reliable identification of skin pixels, particularly for applications that require comprehensive skin coverage. Nevertheless, limitations remain concerning minor precision trade-offs and performance in certain challenging conditions. Subsequent research directions include refining the precision-recall balance without compromising the method's consistent recall performance across diverse datasets.

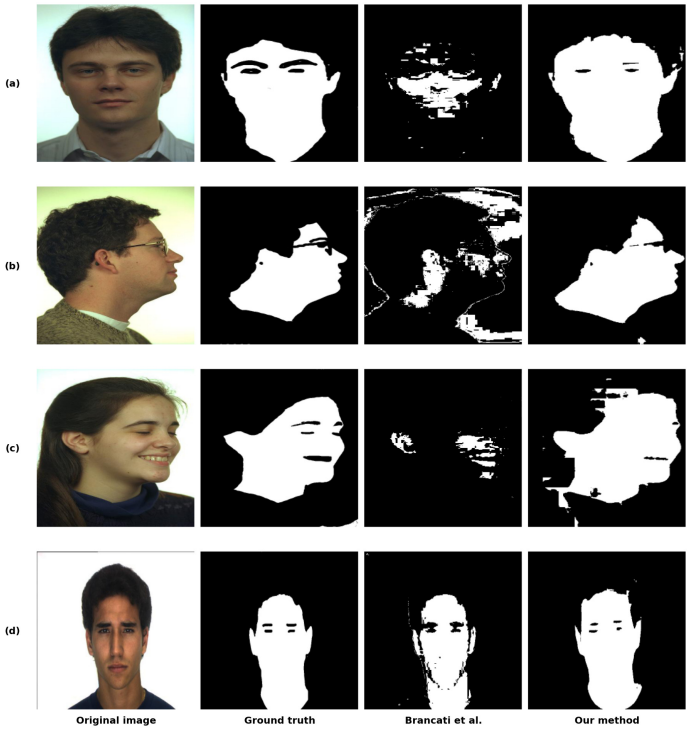


Fig. 9 Qualitative results on the SFA dataset. From left to right: input image, ground truth, [12], proposed method

5 Conclusion

In this paper, we introduced a dual-stream skin detection framework that integrates color cues with CNN-predicted texture descriptors, fused through a hierarchical attention mechanism within a U-Net architecture. By replacing handcrafted texture extraction with a CNN predictor, we significantly reduced computation time while maintaining high accuracy, thereby making the pipeline both efficient and practical. Experimental results across three benchmark datasets — HGR, Pratheepan, and SFA —demonstrated that our method consistently achieves superior recall, reduces false negatives, and yields robust segmentation in complex scenes. Beyond performance improvements, our study underscores the enduring significance of texture when intelligently incorporated into deep learning models. While handcrafted descriptors have limitations, our results suggest that texture remains a valuable cue when modeled in a learned and adaptive manner.

Future scope. In this work, we adopted a fixed value of $\lambda = 0.75$ to balance the contributions of the color and texture streams. While this setting provided satisfactory results in most cases, some scenarios revealed its limitations. A promising direction for future work would be to design a dynamic λ that adapts to image content or prediction confidence, enabling more flexible and context-aware fusion. We also aim to

explore semantic and higher-level representations of texture to enhance further both the interpretability and effectiveness of skin detection systems.
Data availability All datasets used in this work are publically available at: https://sun.aei.polsl.pl/~mkawulok/gestures/ ; http://cs-chan.com/downloads_skin_dataset.html and http://www.sel.eesc.usp.br/sfa/ (accessed 26-september-2021)

Declarations

Competing interests The authors declare that they have no competing interest.

Funding No funding was received for conducting this study.

References

[1] Hsu, R.-L., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color images. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5), 696–706 (2002) https://doi.org/10.1109/34.1000242

[2] Kawulok, M.: Dynamic skin detection in color images for sign language recognition. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) Image and Signal Processing (ICISP), Lecture Notes in Computer Science, Vol. 5099, pp. 112–119. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69905-7_13

[3] Argyros, A.A., Lourakis, M.I.A.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: Pajdla, T., Matas, J. (eds.) Computer Vision – ECCV 2004, Lecture Notes in Computer Science, Vol. 3023, pp. 368–379. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24672-5_29

[4] Zhu, Y., Xu, G., Kriegman, D.J.: A real-time approach to the spotting, representation, and recognition of hand gestures for human–computer interaction. Computer Vision and Image Understanding **85**(3), 189–208 (2002) https://doi.org/10.1006/cviu.2002.0967

[5] Zaidan, A.A., Abdul Karim, H.A., Ahmad, N.N., Zaidan, B.B., Sali, A.: An automated anti-pornography system using a skin detector based on artificial intelligence: A review. International Journal of Pattern Recognition and Artificial Intelligence **27**(4), 1350012 (2013) https://doi.org/10.1142/S0218001413500122

[6] Baykal Kablan, E., Ayas, S.: Skin lesion classification from dermoscopy images using ensemble learning of convnext models. Signal, Image and Video Processing **18**, 6353–6361 (2024) https://doi.org/10.1007/s11760-024-03321-y

[7] Kaur, R., Ranade, S.K.: Du-net+: a fully convolutional neural network architecture for semantic segmentation of skin lesions. Signal, Image and Video Processing **19**, 152 (2025) https://doi.org/10.1007/s11760-024-03690-4

[8] Sakthipriya, G., Padmapriya, N., Venkateswaran, N.: Tsdanet: texture strip dual attention network for intraclass texture classification. SIViP **18**, 7597–7610 (2024) https://doi.org/10.1007/s11760-024-03413-9

[9] Kelishadrokhi, M.K., Ghattaei, M., Fekri-Ershad, S.: Innovative local texture descriptor in joint of human-based color features for content-based image retrieval. SIViP **17**, 4009–4017 (2023) https://doi.org/10.1007/s11760-023-02631-x

[10] Sobottka, K., Pitas, I.: A novel method for automatic face segmentation, facial feature extraction and tracking. Signal Processing: Image Communication **12**(3), 263–281 (1998) https://doi.org/10.1016/S0923-5965(97)00042-8

[11] Kovac, J., Peer, P., Solina, F.: Human skin colour clustering for face detection. In: Proceedings of EUROCON 2003: Computer as a Tool. IEEE Region 8 Conference, vol. 2, pp. 144–148. IEEE, Ljubljana, Slovenia (2003). https://doi.org/10.1109/EURCON.2003.1248169

[12] Brancati, N., De Pietro, G., Frucci, M., Gallo, L.: Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering. Computer Vision and Image Understanding **155**, 33–42 (2017) https://doi.org/10.1016/j.cviu.2016.12.001

[13] Samson, G.L., Lu, J.: Pkt: Fast color-based spatial model for human skin detection. Multimedia Tools and Applications **80**(21–23), 32807–32839 (2021) https://doi.org/10.1007/s11042-021-10955-4

[14] Satheeskumar, R.: Face detection using l*a*b* color space. International Journal for Modern Trends in Science and Technology **8**(2), 105–112 (2022) https://doi.org/10.46501/IJMTST0802017

[15] Wang, X., Zhang, X., Yao, J.: Skin color detection under complex background. In: 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), pp. 1985–1988 (2011). IEEE

[16] Sarkar, A., Abbott, A.L., Doerzaph, Z.: Universal skin detection without color information. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 20–28 (2017). https://doi.org/10.1109/WACV.2017.10

[17] Khan, R., Hanbury, A., Stoettinger, J.: Skin detection: A random forest approach. In: 2010 IEEE International Conference on Image Processing, pp. 4613–4616 (2010). https://doi.org/10.1109/ICIP.2010.5651638

[18] Bush, I.J., Abiyev, R., Ma'aitah, M.K.S., Altıparmak, H.: Integrated artificial intelligence algorithm for skin detection. In: Bardis, N. (ed.) ITM Web of Conferences, Vol. 16 (AMCSE 2017), p. 02004. EDP Sciences, Athens, Greece (2018). https://doi.org/10.1051/itmconf/20181602004

[19] Chakraborty, B.K., Bhuyan, M.K.: Image specific discriminative feature extraction for skin segmentation. Multimedia Tools and Applications **79**, 18981–19004 (2020) https://doi.org/10.1007/s11042-020-08762-4

[20] Tarasiewicz, T., Nalepa, J., Kawulok, M.: Skinny: A lightweight u-net for skin detection and segmentation. In: Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), pp. 2386–2390. IEEE, Abu Dhabi, UAE (2020). https://doi.org/10.1109/ICIP40778.2020.9191209

[21] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science, Vol. 9351, pp. 234–241. Springer, Cham, Switzerland (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[22] Ben Salah, K., Othmani, M., Kherallah, M.: A novel approach for human skin detection using convolutional neural network. The Visual Computer **38**(5), 1833–1843 (2022) https://doi.org/10.1007/s00371-021-02108-3

[23] Leite, M., Parreira, W.D., Fernandes, A.M.d.R., Leithardt, V.R.Q.: Image segmentation for human skin detection. Applied Sciences **12**(23), 12140 (2022) https://doi.org/10.3390/app122312140

[24] Hashemifard, K., Climent-Perez, P., Florez-Revuelta, F.: Weakly supervised human skin segmentation using guidance attention mechanisms. Multimedia Tools and Applications **83**(10), 31177–31194 (2024) https://doi.org/10.1007/s11042-023-16590-5

[25] Khosravi, S.: Chimp optimization algorithm to optimize a convolution neural network for skin detection in hvs and rgb images. International Journal of Nonlinear Analysis and Applications **15**(4), 275–284 (2024) https://doi.org/10.22075/ijnaa.2023.30647.4459

[26] Kuban, P., Kawulok, M.: Ensembling convolutional neural networks for human skin segmentation. In: Correia, L., Rosá, A., Garijo, F. (eds.) Advances in Artificial Intelligence – IBERAMIA 2024, Lecture Notes in Computer Science, Vol. 15277, pp. 185–196. Springer, Cham, Switzerland (2025). https://doi.org/10.1007/978-3-031-80366-6_16

[27] Hu, S., Li, J., Fan, H., *et al.*: A novel adaptive two-stage selection strategy in local binary pattern for texture classification. SIViP **17**, 4039–4048 (2023) https://doi.org/10.1007/s11760-023-02634-8

[28] Temtam, A., Pei, L., Iftekharuddin, K.: Computational Modeling of Deep Multiresolution-Fractal Texture and Its Application to Abnormal Brain Tissue Segmentation (2023). https://arxiv.org/abs/2306.04754

[29] Grzejszczak, T., Kawulok, M., Galuszka, A.: Hand landmarks detection and localization in color images. Multimedia Tools and Applications **75**(23), 16363–16387 (2016) https://doi.org/10.1007/s11042-015-2934-5

[30] Kawulok, M., Kawulok, J., Nalepa, J., Smolka, B.: Self-adaptive algorithm for segmenting skin regions. EURASIP Journal on Advances in Signal Processing **2014**(170), 1–22 (2014) https://doi.org/10.1186/1687-6180-2014-170

[31] Nalepa, J., Kawulok, M.: Fast and accurate hand shape classification. In: Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds.) Beyond Databases, Architectures and Structures. Communications in Computer and Information Science, vol. 424, pp. 364–373 (2014). https://doi.org/10.1007/978-3-319-06932-6_35

[32] Tan, W.R., Chan, C.S., Yogarajah, P., Condell, J.: A fusion approach for efficient human skin detection. IEEE Transactions on Industrial Informatics **8**(1), 138–147 (2012) https://doi.org/10.1109/TII.2011.2172451

[33] Casati, J.P.B., Moraes, D.R., Rodrigues, E.L.L.: Sfa: A human skin image database based on feret and ar facial images. In: Anais do IX Workshop de Visão Computacional. Escola de Engenharia de São Carlos, Universidade de São Paulo, Rio de Janeiro, Brazil (2013). http://iris.sel.eesc.usp.br/wvc/Anais_WVC2013/Poster/2/3.pdf