Article

End-to-End Pixel-Wise Ear Segmentation with U-Net and ResNet-50 Encoder

Mohammad Zahir Uddin Chowdhury * , Avery Shoemaker , Ibrahim Moubarak Nchouwat Ndumgouo , Stephanie Schuckers

Posted Date: 23 April 2026

doi: 10.20944/preprints202604.1693.v1

Keywords: ear biometrics; segmentation; ReNet-50; deep learning

 doi:10.20944/preprints202604.1693.v1

Article

End-to-End Pixel-Wise Ear Segmentation with U-Net and ResNet-50 Encoder

Mohammad Zahir Uddin Chowdhury [1,*], Avery Shoemaker [1], Ibrahim Moubarak Nchouwat Ndumgouo [1] and Stephanie Schuckers [2]

[1] Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY, USA
[2] College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, NC, USA
* Correspondence: mochowd@clarkson.edu

Abstract

Ear biometrics has emerged as a complementary modality for biometric recognition, particularly in unconstrained environments where traditional approaches such as face recognition may be affected by pose, illumination, or occlusion. Accurate ear segmentation plays a critical role in such systems by isolating the region of interest and reducing background interference. However, reliable segmentation remains challenging under real-world conditions due to occlusions, accessories, and variations in image quality. In this work, we investigate an encoder-enhanced U-Net architecture for pixel-wise ear segmentation, incorporating a ResNet-50 backbone to improve feature representation through transfer learning. The proposed approach is evaluated on the Annotated Web Ears (AWE) dataset and the EarSegDB-25 dataset under standard experimental settings. On AWE, the model achieves a mean Intersection over Union (IoU) of 77.1% and a pixel-wise accuracy of 99.7%, outperforming previously reported encoder–decoder baselines. On EarSegDB-25, the method attains a test IoU of 94.76%, demonstrating strong segmentation performance on a dataset with diverse real-world variations. We further analyze the relationship between pixel-wise accuracy and IoU, highlighting the limitations of accuracy as a metric in background-dominated segmentation tasks. While the architectural modification is incremental, the results indicate that incorporating a pretrained residual encoder can provide consistent improvements in segmentation quality under challenging conditions. These findings support the effectiveness of encoder-enhanced U-Net models as a practical solution for ear segmentation in biometric pipelines.

Keywords: ear biometrics; segmentation; ReNet-50; deep learning

1. Introduction

Biometric systems have become indispensable for personal identification and authentication, particularly in domains where security, reliability, and non-intrusiveness are paramount. Among various biometric modalities, the human ear has emerged as a promising trait due to its unique anatomical structure and relative stability with respect to facial expressions, aging, and common accessories such as glasses and facial hair [1]. Ear biometrics are especially valuable in scenarios where facial biometrics may be degraded—such as low-light conditions, non-frontal poses, or partial occlusions—and have been explored for applications including surveillance, forensic analysis, and identity verification. Furthermore, when integrated with facial recognition systems, ear-based recognition can enhance the robustness and accuracy of multi-modal biometric solutions [2].

Despite these advantages, robust and fully automatic ear detection and segmentation in unconstrained environments remains a challenging problem. Common difficulties arise from occlusions (e.g., hair, headwear), pose variations, illumination changes, and differences in image resolution. Moreover, many prior works have relied on constrained or laboratory-captured datasets, which may not fully reflect the variability encountered in real-world applications.

 doi:10.20944/preprints202604.1693.v1

Recent approaches have explored deep learning architectures for ear segmentation, with encoder–decoder networks such as U-Net demonstrating strong potential. However, many implementations employ standard encoders and are trained on relatively limited or homogeneous datasets. For example, Lahkar and Borbora (2024) [3] proposed an automatic ear segmentation system for unconstrained environments, achieving 77.63% IoU on their test set. Another notable study, Convolutional Encoder–Decoder Networks for Pixel-Wise Ear Detection and Segmentation [4], reported an IoU of 55.7%, underscoring the difficulty of attaining high segmentation accuracy under diverse, in-the-wild conditions.

In this work, we evaluate an ear-segmentation approach based on a U-Net architecture with a ResNet50 backbone (Figure 1) using the Annotated Web Ears (AWE) dataset [1]. AWE is a web-harvested collection curated specifically to probe ear recognition under unconstrained settings. In our assessment, and consistent with recent studies [1,5], AWE represents a challenging public benchmark for ear recognition due to its variability in pose, illumination, occlusion, and image quality. Unlike many datasets traditionally used to assess ear detection, AWE is not confined to near-perfect profile faces, but instead spans a broad range of head poses with substantial yaw variation.

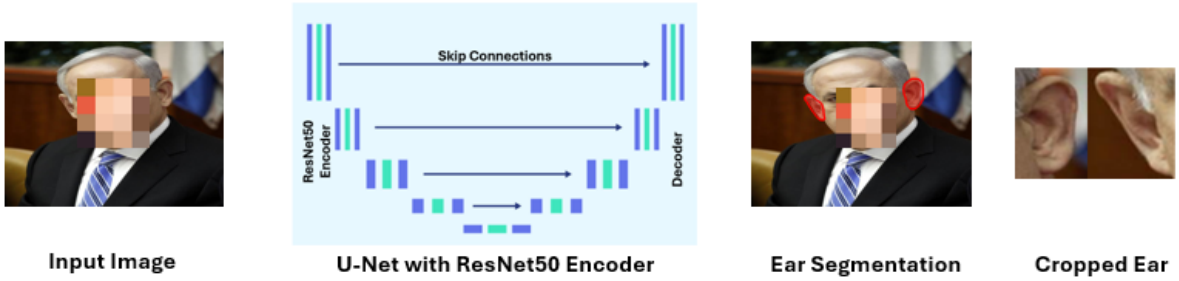

Figure 1. Overview of the proposed ear-segmentation pipeline: a U-Net with a ResNet50 encoder produces a **binary mask** (as shown in red), from which cropped ear regions feed downstream biometric tasks.

Beyond serving as a standalone task, precise ear segmentation is an essential prerequisite for downstream pipelines. For ear recognition, accurate masks and tight region-of-interest extraction directly influence representation quality and matching performance [6]. In this study, we adopt a ResNet50-based U-Net with transfer learning to improve feature representation. The residual encoder leverages ImageNet pretraining, while skip connections preserve spatial detail. Fine-tuning allows adaptation to the ear domain, and data augmentation is used to improve generalization under varying conditions.

Prior studies on the EarSegDB dataset have shown that ear segmentation performance is often limited by low Intersection over Union (IoU), despite reporting relatively high pixel-wise accuracy [3,4]. IoU measures the spatial overlap between the predicted ear region and the ground-truth mask, and low IoU indicates inaccurate localization and poor boundary delineation of the ear, which can adversely affect downstream recognition and analysis tasks. In contrast, pixel-wise accuracy provides only a coarse measure of segmentation quality and is frequently inflated by the dominance of non-ear background pixels in the image. Motivated by these limitations observed in prior work, we emphasize improving IoU as a primary objective, aiming to achieve more accurate ear localization compared to conventional encoder–decoder baselines.

While similar encoder-enhanced U-Net architectures have been explored in other domains—such as medical image segmentation (e.g., tumors, organs) [7,8], remote sensing [9], and autonomous driving [10]—their application to ear segmentation remains relatively limited. These prior studies demonstrate that incorporating pretrained residual encoders can improve feature representation, particularly in settings with limited annotated data. This motivates the use of such architectures for ear segmentation in unconstrained conditions [11].

The main contributions of this work are as follows:

 doi:10.20944/preprints202604.1693.v1

1. An encoder-enhanced U-Net architecture based on a ResNet50 backbone for pixel-wise ear segmentation.
2. Empirical evaluation on AWE demonstrating improved IoU and pixel accuracy compared to prior encoder–decoder baselines [3,4].
3. Release of trained models and a cropping pipeline for reproducibility (GitHub Repository).
4. Analysis of metric behavior, highlighting the limitations of pixel-wise accuracy in background-dominated segmentation tasks.
5. Comparative evaluation under varying pose, occlusion, and illumination conditions [1,5].

This work provides an empirical assessment of an encoder-enhanced U-Net model for ear segmentation, focusing on performance under unconstrained conditions and practical applicability in biometric pipelines. The study emphasizes empirical evaluation and analysis rather than proposing a fundamentally new segmentation architecture.

2. Literature Review

This section presents a comprehensive review of datasets and methodologies used in ear biometrics, with a focus on detection and segmentation techniques. The discussion begins with datasets commonly adopted for evaluation, followed by a categorization of existing methods based on their technical approach.

2.1. Ear Biometrics Datasets

A variety of datasets have been developed to support research in ear detection and recognition. These include classic controlled datasets like USTB, UND, and IIT Delhi, as well as more recent large-scale, unconstrained collections such as VGGFace-Ear, HelloEars, EarVN1.0, and AWE. Some datasets emphasize high-resolution imaging (e.g., UBEAR, AWE), while others prioritize variations in pose, illumination, and occlusion.

Among the unconstrained datasets, we selected the AWE dataset for our experiments due to its moderate size, public availability, and challenging in-the-wild conditions, including diverse subjects, occlusions, and varying head poses. Other unconstrained datasets were not suitable for our purposes for the following reasons: *VGGFace-Ear* is not publicly available for download, *HelloEars* lacks consistent annotations and standardized evaluation protocols, *EarVN1.0* is relatively small in size and resolution, and *UERC* is primarily designed for recognition benchmarking with limited segmentation ground truth. Therefore, AWE provides an optimal balance between accessibility, diversity, and data quality for training and evaluating ear detection and segmentation systems. These datasets are summarized in Table 1, 2.

Table 1. Summary of constrained ear image datasets.

Dataset	Year	Subjects	Images
USTB-1 [12]	2002	60	180
USTB-1 [12]	2003	77	308
USTB-1 [12]	2004	79	1738
USTB-1 [12]	2007	500	8500
UND_E [13]	2002	302	942
UND_F [13]	2003	114	464
UND_G [13]	2005	235	738
UND_E-J2 [13]	2005	415	1800
IITD_1 [14]	2007	121	471
IITD_2 [14]	2008	212	754
IIITK_1 [15]	2019	119	471
IIITK_2 [15]	2019	89	754

 doi:10.20944/preprints202604.1693.v1

Table 2. Summary of unconstrained ear image datasets.

Dataset	Year	Subjects	Images
AWE [1]	2017	100	1000
VGGFace-Ear [16]	2022	660	234651
EarVN1.0 [17]	2018	164	28412
HelloEars [18]	2017	1570	610000
UBEAR [19]	2011	126	4330
AMI [20]	2008	100	700
UERC [21]	2019	3690	11000
WPUT [22]	2010	501	2071
OP1B [23]	2023	152	907
EarSegDB [24]	2023	24	1275

2.2. Methodologies for Ear Detection and Segmentation

Ear detection and segmentaion algorithms can broadly be categorized into seven methodological classes, reflecting the progression from semi-automated rule-based systems to modern deep learning models.

Semi-Automated Methods: Semi-automated methods rely on partial user input, geometric modeling, or initialization assumptions. Yan and Bowyer [25] introduced the Two-Line Landmark technique, which requires manual placement of landmarks. Alvarez et al. [26] proposed the Ovoid Model to fit an elliptical contour over the ear's outer edge. Deepak et al. [27] presented a two-stage Active Contour Model using background removal and SVM-based feedback for ear localization.

These approaches perform well in controlled settings but may degrade under occlusion, pose variation, or background complexity.

Template Matching: Template matching identifies ears by comparing image features to predefined templates. Ansari and Gupta [28] used edge maps and convex curve detection to match ear contours. Surya Prakash and Gupta [29] improved robustness by using distance-transformed templates. Additional methods utilize connected components [30] or histogram-based 2-phase matching [31].

While these techniques yield high accuracy on constrained datasets like IITK, they remain sensitive to occlusions or image noise.

Morphological Operators: Morphological methods transform pixel structures to isolate ear regions. Said et al. [32] introduced a four-stage approach: top-hat filtering, k-means clustering, connected component analysis, and post-processing. These techniques are efficient and interpretable, but may fail under background clutter or poor lighting.

Shape-Based Methods: These methods exploit the geometric consistency of the ear. The Hough Transform [15] detects circular arcs and is robust to missing or broken edges. The Ray Transform [33] emphasizes elliptical features by simulating the propagation of rays across the image.

Both methods achieve high accuracy under varying poses and illumination.

Wavelet-Based Methods: Wavelet-based methods perform multi-resolution analysis. Ibrahim et al. [34] introduced the Banana Wavelet approach using curved wavelets tuned to the ear's contour. It demonstrated 100% detection accuracy on the XM2VTS dataset and showed strong performance in noisy or occluded environments.

Learning-Based Methods: Learning-based approaches use machine learning classifiers or rule-based logic with handcrafted features. Abdel Wahab et al. [35] proposed HEARD, a geometric rule-based model using edge characteristics like roundness and area-to-perimeter ratio. It achieved 98% detection accuracy on the UND dataset with efficient runtime and good robustness to occlusion.

Deep Learning Methods: Deep learning methods dominate recent literature due to their ability to learn rich features directly from data. SegNet [4], YOLOv2 [36], and Faster R-CNN [37] have achieved high detection accuracy in both controlled and wild settings. Zhang et al. [37] proposed a multi-scale Faster R-CNN to detect ear, pan-ear, and head regions jointly. Zhang et al. [38] further extended this to 3D detection and normalization using Faster R-CNN + LSV filtering + ICP-based alignment, achieving 100% recognition on CASIA and 98.55% on UND-J2.

All the methods described above, including their respective categories, datasets, accuracy scores, and technical characteristics, are summarized in the comprehensive Table 3.

 doi:10.20944/preprints202604.1693.v1

Table 3. Comprehensive Comparison of Ear Detection & Segmentation Methods.

Category	Method	Database	Acc. (%)	Remarks
Semi-Automated	Two-line Landmark [25]	UND-f	84.1	Not robust to pose or hair occlusion
	Modified Snake + Ovoid Model [26]	NA	98	Semi-automated model
	Active Contour Model [27]	UND	85.5	High computational time
Template Matching	Canny edge detector [28]	IITK	93.34	Outer helix curve based
	Distance transform [29]	IITK	95.2	Sensitive to quality and occlusion
	Edge map + Connected Component [30]	IITK	99.25	High accuracy
	2-Phase Template Match [31]	3D Range Images	91.5	Histogram-based detection
Morphological	Fourier Descriptor [14]	IIT Delhi	NA	Used Fourier descriptors
	Top-hat + KMeans [32]	Various	> 90	Region labeling and refinement
Shape-Based	Hough Transform [15]	UND	91	Performance drops with occlusion
	Ray Transform [33]	XM2VTS	99.6	Robust to shape, illumination
Wavelet-Based	Banana wavelets [34]	XM2VTS	100	Robust to high noise
Learning-Based	HEARD [35]	UND	98	Used geometric thresholds
Deep Learning	Faster R-CNN [37]	WebEar, UND J2, UBEAR	98–100	Multi-scale deep learning
	YOLOv2 [36]	USTB, WebEar, HelloEar	96.17, 98.67	Real-time detection
	CNN + Geometric Morphometrics	CVL	95	Robust to partial occlusion
	Faster R-CNN + Spatial Attention [38]	UND-J2, UBEAR	98–100	Combines region proposal
	SegNet [4]	AWE	99.21	Pixel-wise segmentation

3. Methodology

3.1. Datasets Used

3.1.1. AWE Dataset

The Annotated Web Ears (AWE) dataset [1] is a comprehensive benchmark for ear biometrics. It contains 1,000 web-scraped images with large variation in pose, illumination, resolution, and occlusion. We randomly selected 750 images for training and 250 for testing. Training images were used to learn the parameters of the U-Net segmentation model; the held-out test set was used only for final evaluation. Figure 2 presents several randomly chosen images together with their corresponding masks. The annotations capture ear locations with greater detail than simple rectangular boxes and have been employed in earlier studies [39,40].

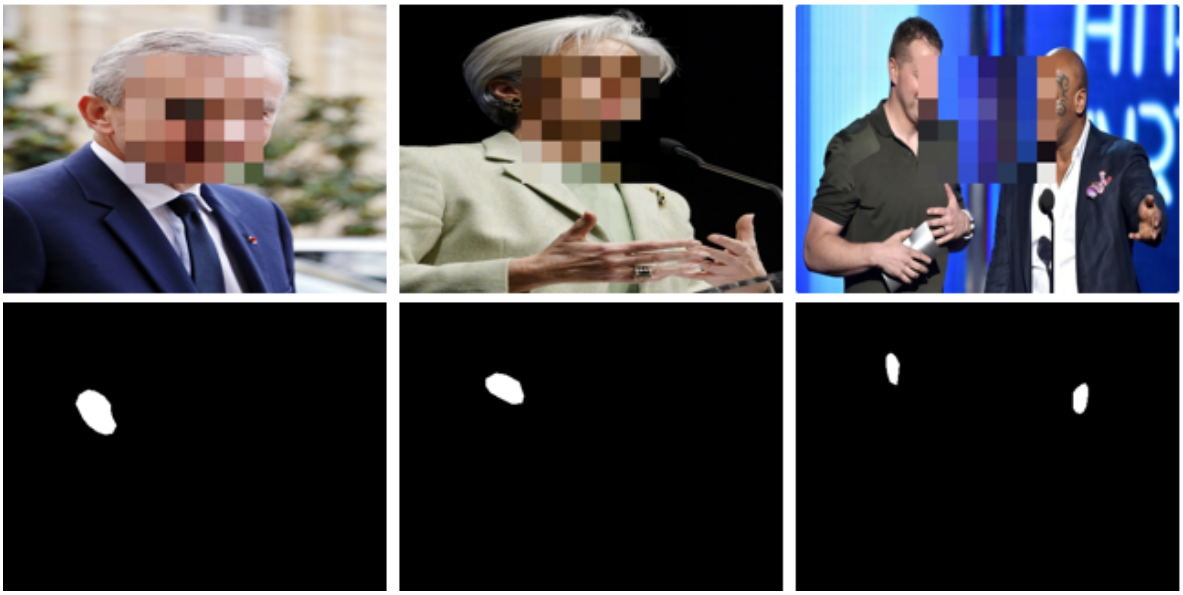

Figure 2. Random samples from the AWE segmentation dataset with their corresponding masks (faces pixelated for anonymity)[1].

3.1.2. EarSegDB 25 Dataset

EarSegDB 25 [24] comprises 1,275 ear images from 25 subjects, each paired with a binary pixel-wise ground-truth mask. The images were captured using smartphone cameras in unconstrained ("in-the-wild") conditions, exhibiting substantial variations in lighting, pose, and viewpoint. The dataset also provides subject-level metadata, including age, gender, and ear side (left/right). In this work, we use EarSegDB 25 following the original dataset protocol, employing the provided training split for model learning and the official test split for evaluation, without any additional re-splitting. The availability of manually annotated pixel-level masks enables objective benchmarking of ear

 doi:10.20944/preprints202604.1693.v1

segmentation performance using region-based metrics such as Intersection over Union (IoU), and pixel accuracy. As one of the largest publicly available ear datasets with pixel-wise annotations, EarSegDB 25 addresses a critical gap in ear biometrics research, where the lack of annotated segmentation ground truth has historically limited quantitative and reproducible evaluation of ear segmentation methods. Example images and corresponding masks are shown in Figure 3.

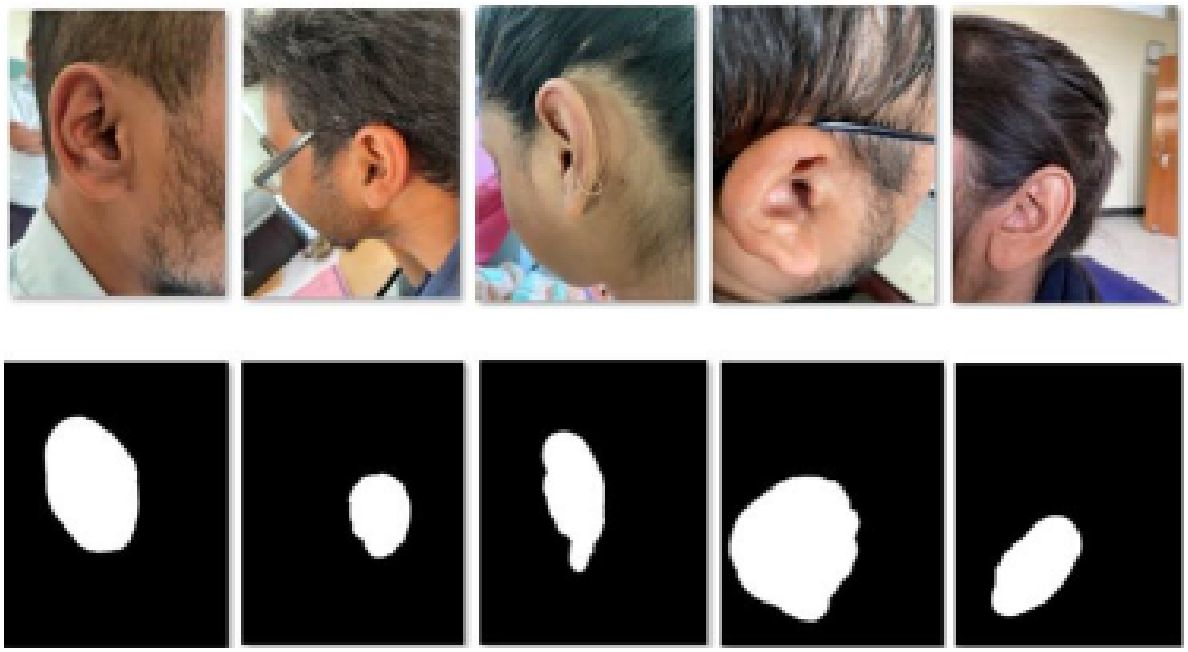

Figure 3. Random images from the EarSegDB 25 dataset with respective masks [3].

This study uses publicly available datasets (AWE and EarSegDB-25), and no new human subject data were collected. Ethical approval and participant consent were handled by the original dataset creators.

3.2. U-Net Architecture

U-Net was introduced by Ronneberger *et al.* for biomedical image segmentation as a fully convolutional network that performs well even with limited annotated data [41]. The model derives its name from its characteristic *"U"-shaped* topology: a left branch that compresses the image into compact feature maps and a right branch that expands them back to full resolution (see Figure 4). This shape reflects the design goal—first aggregate broad context, then restore fine spatial detail for dense, pixel-wise prediction.

 doi:10.20944/preprints202604.1693.v1

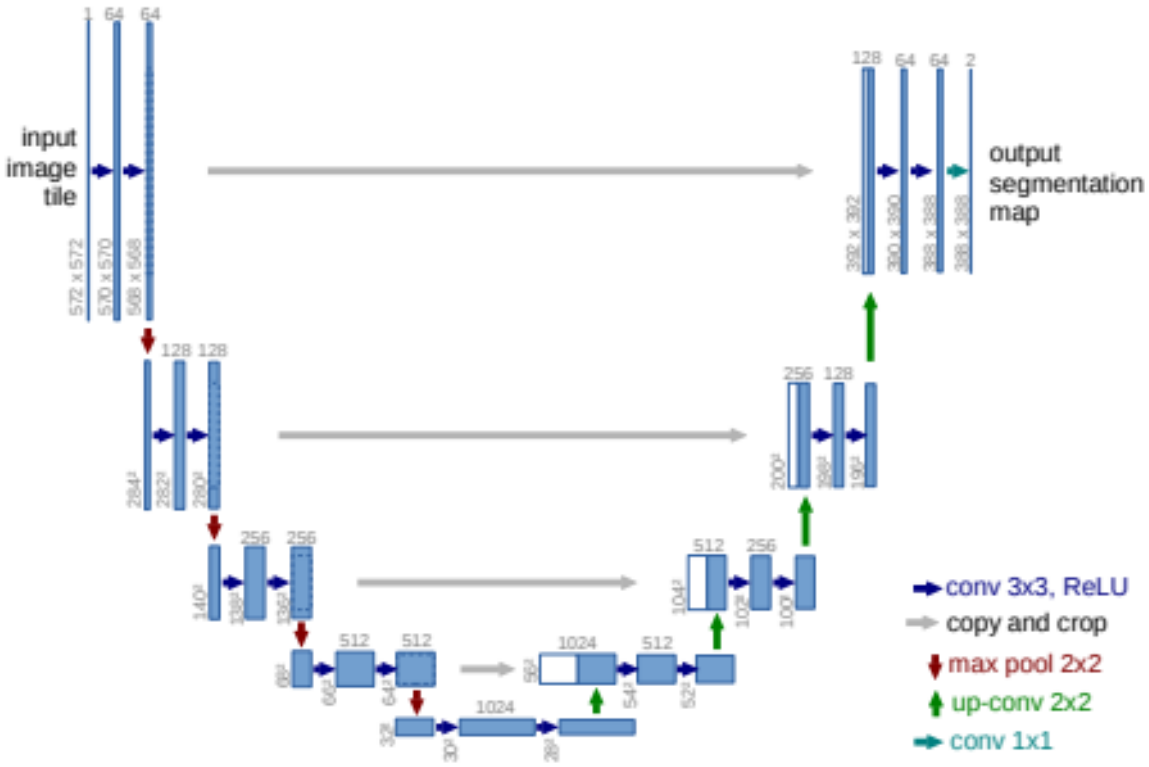

Figure 4. The U-Net Architecture [41].

The *contracting path* (encoder) repeatedly applies two 3×3 convolutions with ReLU followed by 2×2 max pooling. Each pooling step halves the spatial resolution while typically doubling the number of channels (e.g., $64 \to 128 \to 256 \to 512 \to 1024$). The *expanding path* (decoder) upsamples with 2×2 transposed convolutions, concatenates the matching high-resolution features from the encoder via *skip connections*, and refines them with 3×3 convolutions. In the original implementation, convolutions used *valid* (no-padding) kernels, so encoder feature maps are cropped before concatenation to align sizes; many modern variants adopt "same" padding to avoid cropping while retaining the same logic [41].

Beyond the layout, U-Net's training and inference choices are crucial for strong results with scarce labels. The original work employs *strong data augmentation*—notably elastic deformations, plus rotations, shifts, and intensity changes—to simulate variability. Optimization is performed with a pixel-wise softmax cross-entropy and a *boundary-weighted* loss map that helps separate touching objects. At test time, an *overlap-tile* strategy with mirrored borders enables seamless segmentation of large images without edge artifacts [41]. These practices, together with skip connections, explain how U-Net maintains sharp boundaries while leveraging global context.

For ear segmentation, these properties are directly beneficial. Ears contain thin, high-frequency contours that demand precise localization, yet their appearance varies with pose, illumination, hair, and accessories. U-Net's encoder aggregates the necessary global context to disambiguate ear versus background, while the decoder—fed by skip connections—preserves fine boundary detail. With appropriate augmentation and a boundary-aware loss, the network converges reliably and yields accurate masks even from relatively modest datasets, consistent with the strengths reported in the original U-Net paper [41].

3.3. Proposed Model: U-Net with ResNet50 Encoder

In our approach, we integrate a pre-trained ResNet50 network as the encoder part of the U-Net (see Figure 5). The ResNet50 model, known for its residual learning framework [42], significantly improves feature extraction by enabling deeper networks without vanishing gradient problems. The encoder consists of convolutional blocks with residual connections, capturing high-level features

 doi:10.20944/preprints202604.1693.v1

efficiently. The decoder consists of transposed convolutional layers that progressively reconstruct the segmentation mask. Skip connections from the encoder to the decoder help recover fine-grained spatial details by combining low-level and high-level features. The final layer uses a sigmoid activation function to generate the binary segmentation mask. Architecture breakdown is given in Table 4:

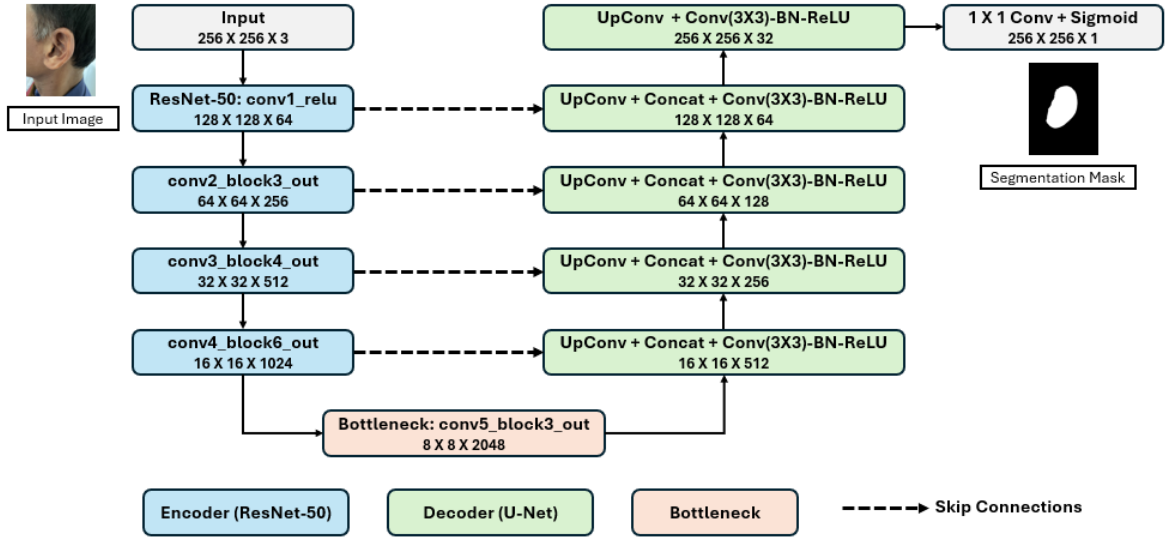

Figure 5. The modified U-Net: a trainable ResNet-50 encoder on the left (skip taps at conv1/2/3/4), a bottleneck at the bottom, and a U-Net decoder on the right using UpConv + Conv(3×3)–BN–ReLU blocks. Dashed arcs denote skip concatenations; a final 1×1 sigmoid outputs the binary mask.

Table 4. U-Net with ResNet50 Encoder Architecture.

Stage	Operation	Output Size	Filter Size
Input	Input RGB Image	$256 \times 256 \times 3$	–
Encoder	Conv1 + ReLU + MaxPool	$128 \times 128 \times 64$	7×7 conv, 3×3 pool
Encoder	Residual Block	$64 \times 64 \times 256$	3×3 conv
Encoder	Residual Block	$32 \times 32 \times 512$	3×3 conv
Encoder	Residual Block	$16 \times 16 \times 1024$	3×3 conv
Bottleneck	Residual Block	$8 \times 8 \times 2048$	3×3 conv
Decoder	UpConv + Skip	$16 \times 16 \times 512$	2×2 upconv
Decoder	UpConv + Skip	$32 \times 32 \times 256$	2×2 upconv
Decoder	UpConv + Skip	$64 \times 64 \times 128$	2×2 upconv
Decoder	UpConv + Skip	$128 \times 128 \times 64$	2×2 upconv
Decoder	UpConv	$256 \times 256 \times 32$	2×2 upconv
Output	1×1 Conv + Sigmoid	$256 \times 256 \times 1$	1×1 conv

3.4. Implementation Summary

Framework. All experiments were run in TensorFlow/Keras with Albumentations for data augmentation. Images are resized to 256×256 RGB and masks to 256×256 grayscale; intensities are normalized to $[0, 1]$ and masks are binarized at 0.5.

Data split and augmentation. For each dataset we load pairs (x, m) from disk and discard samples with missing masks. We follow dataset-specific evaluation protocols. For the AWE dataset, we use a fixed split of 750 images for training and 250 for testing, consistent with prior studies. For EarSegDB-25, we adopt the official dataset-defined training and test splits. Only the training data is augmented using a fixed pipeline: horizontal flip, random brightness/contrast, shift–scale–rotate, random resized crop, elastic transform, and grid distortion. The augmented images are used to train the model in place of the original training images.

 doi:10.20944/preprints202604.1693.v1

3.5. Network Architecture

Encoder (ResNet-50). We adopt the Keras `ResNet50` backbone with `include_top=False` and `weights="imagenet"`. We tap four skip connections at `conv1_relu` (128×128), `conv2_block3_out` (64×64), `conv3_block4_out` (32×32), and `conv4_block6_out` (16×16); the bottleneck is `conv5_block3_out` (8×8). *All encoder layers remain trainable* (no layer freezing), so the network is fine-tuned end-to-end from ImageNet initialization.

Decoder (U-Net style). The decoder mirrors the encoder with 2×2 transposed convolutions (stride 2) to upsample to 16×16, 32×32, 64×64, 128×128, and 256×256. At each scale we concatenate the corresponding encoder skip feature and apply a *conv block* of Conv(3×3)–BatchNorm–ReLU with channel sizes $\{512, 256, 128, 64, 32\}$. A final 1×1 convolution with σ produces a single-channel probability map.

3.6. Training Setup

Hyperparameters. Images: 256×256. Optimizer: Adam with learning rate 10^{-3}. Batch size: 2 (GPU-memory friendly). Epoch budget: 100. Schedulers: `ReduceLROnPlateau` (monitor `val_loss`, factor 0.5, patience 5) and `EarlyStopping` (patience 10, `restore_best_weights=True`). We checkpoint the best model by validation loss.

3.7. Departures From the Original U-Net [41]

- **Backbone replacement.** The original U-Net encoder is a shallow stack of conv–pool blocks. We replace it with a **pretrained ResNet-50** and expose four hierarchical skip taps (`conv1`, `conv2`, `conv3`, `conv4`). This deep, residual encoder improves feature richness and convergence. *No layers are frozen*; the encoder is fine-tuned end-to-end.
- **Padding and alignment.** U-Net (2015) used *valid* convolutions with cropping before skip concatenation. Our implementation uses Keras layers with *same* padding throughout the decoder (and in ResNet), removing the need for cropping and keeping the spatial size at 256×256.
- **Decoder details.** We use 2×2 transposed convolutions for upsampling followed by Conv(3×3)–BatchNorm–ReLU blocks with channel sizes $\{512, 256, 128, 64, 32\}$; the original used pairs of 3×3 convolutions without BatchNorm.
- **Loss and optimization.** Instead of a boundary-weighted softmax cross-entropy, we employ **BCE** with **Adam** (10^{-3}). Learning-rate scheduling and early stopping are used for stable convergence.
- **Augmentation.** We extend the augmentation recipe with brightness/contrast jitter, random resized crop, elastic transform, grid distortion, and shift–scale–rotate. These augmentations target pose/illumination variability typical of ear images.
- **Input resolution and masks.** We train on fixed 256×256 images (original U-Net used larger valid-padded tiles) and binarize masks at 0.5.

3.8. Training Protocol and Loss Convergence

We trained separate models on the two datasets using standard preprocessing and train/validation splits for robust generalization.

The loss curves in Figure 6 show convergence over 22 epochs:

- **Blue** = training loss; **orange** = validation loss.
- **Initial phase (epochs 0–2):** training loss starts around 0.1 and falls quickly; validation loss begins very high (peak ≈ 0.9) then drops steeply—typical of unstable early learning.
- **Rapid learning (epochs 2–4):** both curves decline sharply as the model starts capturing salient features.
- **Adjustment (epochs 4–8):** validation shows small oscillations (brief bumps) while training decreases more smoothly—expected due to batch variation and unseen validation data.
- **Convergence (∼epoch 10 onward):** both losses flatten close to zero with a small, stable gap (validation slightly above training), indicating no overfitting.

 doi:10.20944/preprints202604.1693.v1

- **Outcome:** very low, aligned losses suggest **excellent generalization** and **robust segmentation performance** on the validation set; early stopping around epochs 10–12 would likely retain performance.

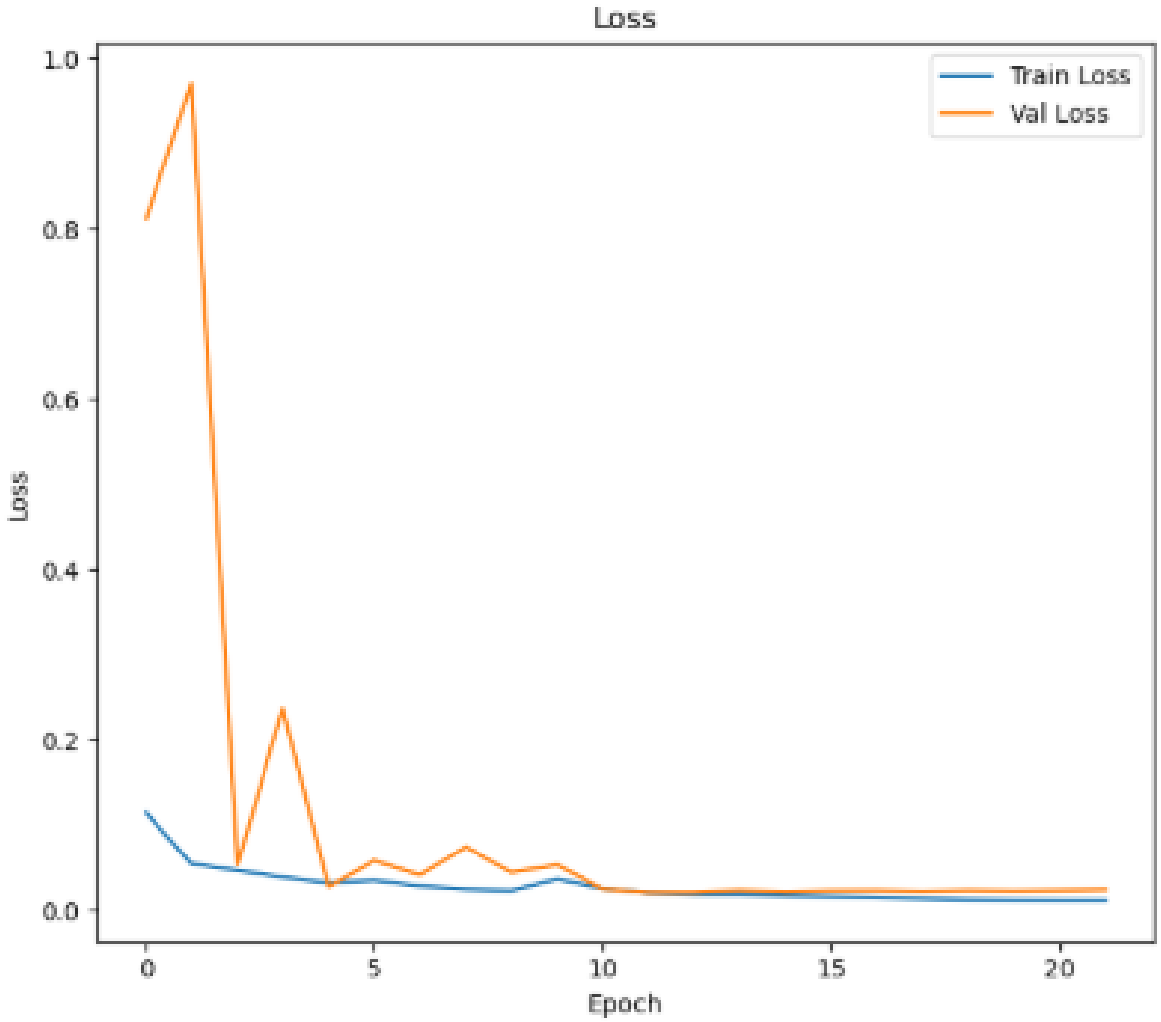

Figure 6. Training and Validation Loss Curve.

3.9. Feature Extraction Visualization with ResNet50

To encode images for segmentation, we employ a *ResNet50* backbone that learns hierarchical features through residual learning. Identity shortcuts (skip connections) ease optimization in very deep networks by mitigating vanishing gradients, enabling the encoder to extract progressively more abstract representations. Early stages emphasize low-level edges and textures; intermediate stages aggregate contours and parts; deeper stages become spatially sparse yet semantically rich, focusing on the most informative regions for segmentation.

Figure 7 visualizes feature maps sampled from five depths of the encoder, illustrating this progression:

- `conv1_relu`: Basic edge/texture responses and broad shapes.
- `conv2_block3_out`: Clearer local patterns; emerging ear/face contours.
- `conv3_block4_out`: Stronger abstractions; noise suppression and emphasis on salient structures.
- `conv4_block6_out`: Spatially sparser, semantically richer activations focused near ear boundaries.
- `conv5_block3_out`: Very sparse, highly discriminative features that guide mask prediction.

 doi:10.20944/preprints202604.1693.v1

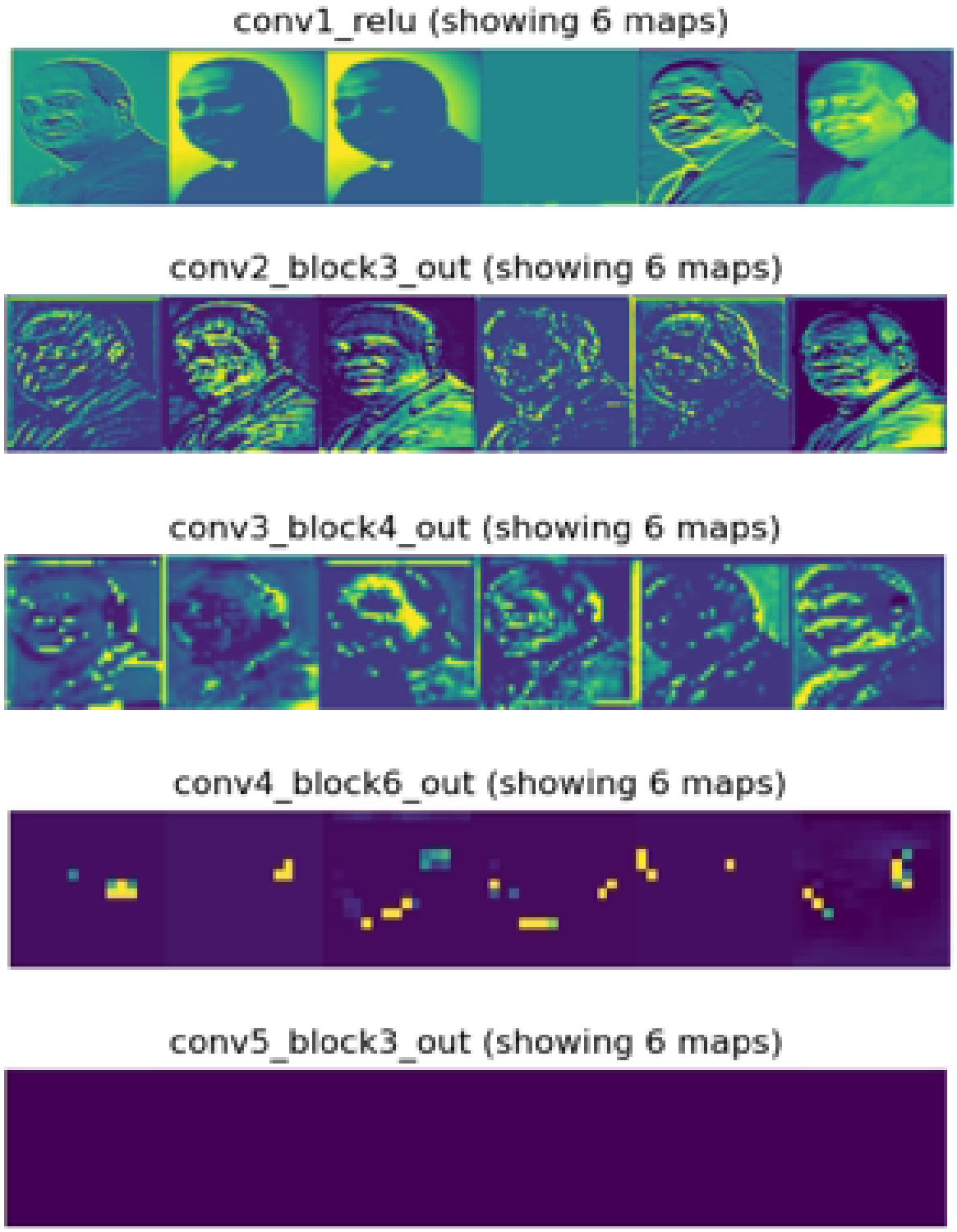

Figure 7. Feature Maps at Different Layers of ResNet50.

These multi-scale features are routed via U-Net skip connections (from `conv1`/`conv2`/`conv3`/`conv4`) to the decoder, allowing precise boundary recovery while preserving high-level semantic context.

3.10. Performance Metrics

In this work, we evaluate the performance of our detection method using the same metrics as described in [4]. Specifically, we compare the manually annotated ground-truth ear masks with the predicted outputs generated by our U-Net approach on the test dataset.

 doi:10.20944/preprints202604.1693.v1

We report several performance metrics, starting with **Accuracy**, defined as

$$\text{Accuracy} = \frac{TP + TN}{\text{All}}, \tag{1}$$

where $\text{All} = TP + TN + FP + FN$. Here:

- **TP (True Positives)**: ear pixels correctly identified as ear pixels,
- **TN (True Negatives)**: non-ear pixels correctly identified as non-ear pixels.

While Accuracy provides a basic indication of segmentation quality, it can be skewed due to the large proportion of non-ear pixels that dominate most images. Consequently, even a high Accuracy value (close to 1) may not fully reflect the model's performance on ear regions.

This limitation is illustrated in Figure 8. In the example, only 50% of the ear pixels are correctly segmented by the predicted mask. However, due to the dominance of background pixels, the pixel-wise Accuracy reaches 90%, which may incorrectly suggest strong segmentation performance. In contrast, the corresponding IoU value is 50%, accurately reflecting the partial overlap between the predicted and ground-truth ear regions.

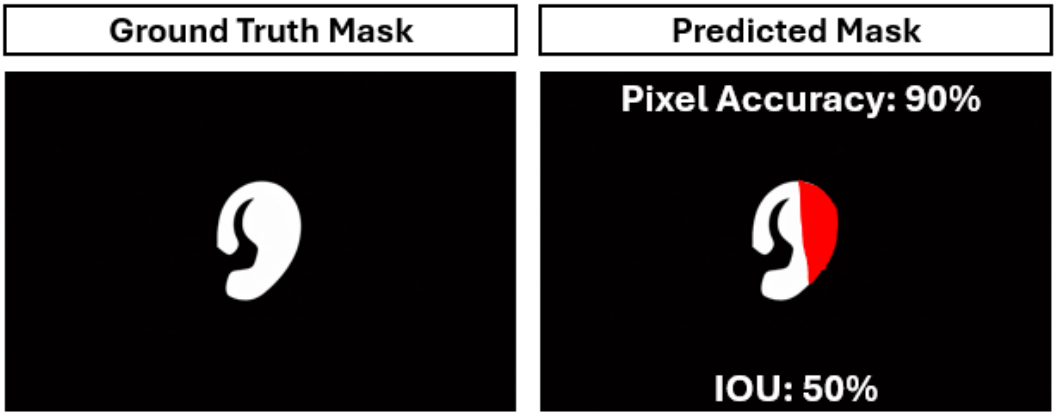

Figure 8. Illustrative comparison between pixel-wise Accuracy and Intersection over Union (IoU) for segmentation evaluation. Although the predicted mask correctly segments only 50% of the ear region, the pixel-wise Accuracy reaches 90% due to the dominance of background pixels. In contrast, IoU correctly reflects the true overlap between the predicted and ground-truth ear regions, yielding a value of 50%.

The second metric employed is the **Intersection over Union (IoU)**, computed as

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{2}$$

where:

- **FP (False Positives)** are non-ear pixels incorrectly classified as ear pixels,
- **FN (False Negatives)** are ear pixels incorrectly classified as non-ear pixels.

IoU measures the overlap between the predicted and ground-truth ear regions. An IoU of 1 indicates perfect alignment; an IoU of 0 implies no overlap. For this reason, IoU is treated as the primary evaluation metric in this work.

We also report **Precision** and **Recall**:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}. \end{aligned} \tag{3}$$

 doi:10.20944/preprints202604.1693.v1

Precision assesses the proportion of true ear pixels among all pixels predicted as ear, reflecting the relevance of detections; Recall evaluates the proportion of true ear pixels successfully detected by the model, indicating detection completeness.

Lastly, we use the E_2 **Error**, which is especially useful in scenarios with class imbalance between ear and non-ear pixels. It is calculated as the average of the False Positive Rate (FPR) and False Negative Rate (FNR):

$$\begin{aligned} E_2 &= \frac{\text{FPR} + \text{FNR}}{2}, \\ \text{FPR} &= \frac{FP}{TN + FP}, \\ \text{FNR} &= \frac{FN}{FN + TP}. \end{aligned} \tag{4}$$

A lower E_2 value signifies better model performance, with $E_2 = 0$ representing perfect precision and recall (no false positives and no false negatives).

4. Results and Analysis

We evaluate our the U-Net model independently on two benchmarks. Evaluation on EarSegDB-25 presents representative segmentations and a detailed metric breakdown. Evaluation on AWE reports the corresponding analysis for the AWE dataset.

4.1. Evaluation on EarSegDB-25

We benchmark our U-Net with a ResNet-50 encoder against the recent method of Lahkar *et al.* [3] for automatic ear segmentation in unconstrained conditions. We focus on this comparison primarily because Lahkar *et al.* used the same dataset (EarSegDB-25) with identical train/validation/test splits and evaluation metrics, which allows for a direct apples-to-apples comparison. To the best of our knowledge, no other peer-reviewed studies have reported segmentation performance on this specific dataset to date. For a direct, pixel-level comparison, both approaches are evaluated using *Accuracy* and *Intersection over Union (IoU)*.

Table 5 summarizes results across the train/validation/test splits. Our model consistently outperforms the prior work on all splits and metrics. On the **training** set, we observe higher Accuracy and IoU compared to [3]. On the **validation** set, our method again achieves better Accuracy and IoU than the baseline. The largest margins appear on the **test** set, where our method reaches **98.72%** Accuracy and **94.76%** IoU, surpassing the results reported by [3].

Table 5. Comparative performance against Lahkar *et al.* [3] on EarSegDB25 dataset.

Split	Metric	L&B [3]	Ours
Training	Accuracy (%)	93.98	96.36
	IoU (%)	83.37	93.00
Validation	Accuracy (%)	92.98	96.22
	IoU (%)	79.33	92.74
Test	Accuracy (%)	92.38	98.72
	IoU (%)	77.63	94.76

Interestingly, our test IoU (**94.76%**) is marginally higher than the training (**93.00%**) and validation (**92.74%**) IoUs. While less common, this effect is not problematic; it suggests that the test distribution is slightly less challenging or is better aligned with the learned feature space. Similar behavior has been noted in prior segmentation literature (e.g., U-Net [41] and UNet+[43]), where evaluation subsets with lower variability or cleaner boundaries yield slightly higher scores than training/validation segments. Taken together, the consistent gains across all splits and the strong test performance indicate robust generalization in unconstrained ear imagery.

 doi:10.20944/preprints202604.1693.v1

Figure 9 presents qualitative examples from the test set. In each panel, the red region denotes the predicted ear mask and the green contour traces the ground-truth boundary. The IoU for each image is annotated in the upper-left corner.

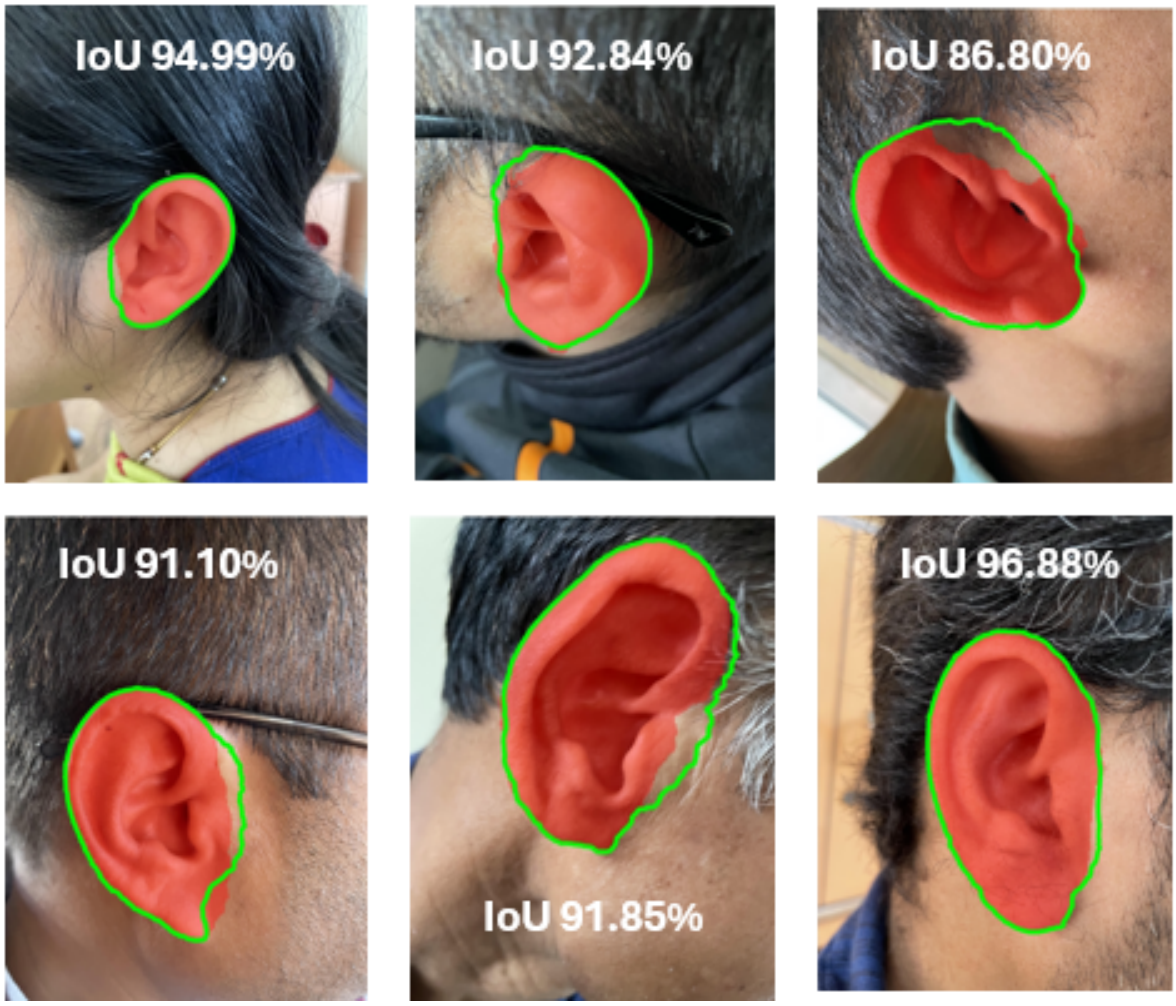

Figure 9. Qualitative ear segmentation results on the EarSegDB-25 test set. Predicted masks are shown in red and ground-truth boundaries in green. Each panel reports the per-image IoU. High IoU values reflect tight alignment of the predicted mask with the annotation; lower IoU indicates minor boundary deviations, often near hair or eyeglass occlusions.

Overall, the predictions adhere closely to the annotated ear boundaries, with IoUs ranging from **86.8%** to **96.9%** in the examples shown. Higher IoUs (e.g., 94.99%, 96.88%) indicate near-perfect overlap between the predicted mask and the ground truth. The case with **86.80%** IoU shows a modest boundary discrepancy, typically arising near hair, glasses frames, or low-contrast edges—common challenges in unconstrained imagery. These visual results complement the quantitative metrics by illustrating that the model captures the full pinna structure while maintaining accurate boundaries across different subjects and viewpoints.

Figure 10 shows an example where the model performs poorly (**IoU = 41.68%**). The predicted mask (red) spills into the skin and misses parts of the ear inside the green ground-truth outline. This likely happens because the ear boundary is hard to see (low contrast), there is shine on the skin, or hair/background colors are similar to the ear. To reduce such errors, we plan to train with more images that include hair and lighting variations, strengthen edge cues, and apply a light post-processing step to clean the mask.

 doi:10.20944/preprints202604.1693.v1

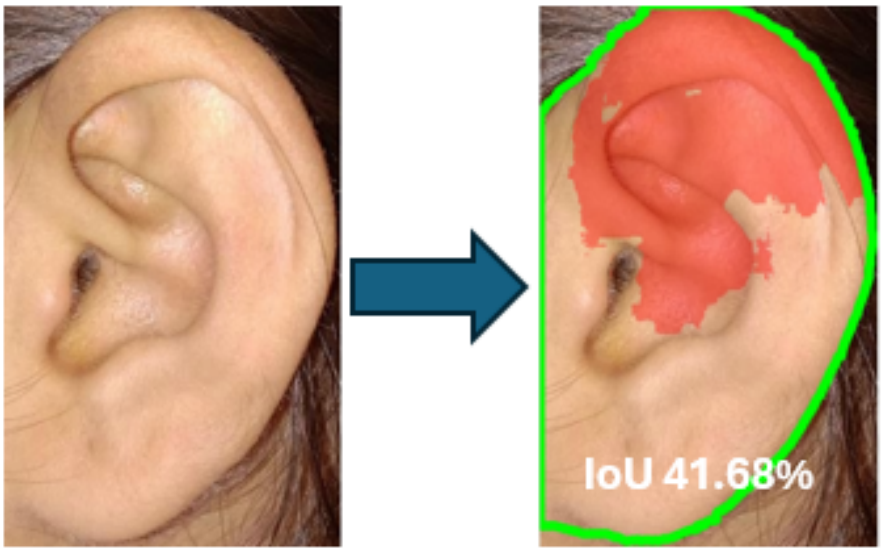

Figure 10. Challenging case with low overlap (IoU = 41.68%). Left: original image. Right: predicted mask (red) with ground-truth boundary (green). The model over-segments into surrounding skin and under-segments parts of the pinna, yielding a low IoU. Such errors typically arise from hair/skin color similarity, specular highlights, and weak edges around the helix/antihelix in unconstrained settings.

4.2. Evaluation on AWE

We compare our U -Net (ResNet-50 encoder) against the baselines reported in [4], which evaluated multiple algorithms including Haar cascades, SegNet, and PED-CED, thereby enabling a direct side-by-side comparison. Metrics include pixel-wise **Accuracy**, **Intersection over Union (IoU)**, **Precision**, **Recall**, and the E_2 **error** (lower is better; see Sec. 3.10 for definitions). Baseline numbers are taken from [4], and our results are computed under the same evaluation protocol. We specifically chose to compare with Emersic *et al.* to allow for an apples-to-apples evaluation, as their study defined fixed train/test splits on the AWE dataset and reported all five key metrics along with standard deviations (±SD), making direct comparison meaningful. We further extend this evaluation by including precision–recall (PR) curves and inference time analysis to provide a deeper assessment. In addition to our ResNet-based model, we also evaluated a plain U-Net architecture (without ResNet50 encoder) for a fair comparison. As shown in Table 6, this provides further insight into the performance gains from using a deeper encoder and transfer learning

Table 6. Comparison of segmentation performance with methods from [4]. Values are mean ± std (%); ↑ higher is better, ↓ lower is better.

Method	Accuracy, % ↑	IoU, % ↑	Precision, % ↑	Recall, % ↑	E_2 ↓
Haar [4]	98.8 ± 1.1	27.2 ± 36.5	36.7 ± 46.6	28.5 ± 38.4	36.4 ± 18.2
SegNet [4]	99.2 ± 0.6	48.3 ± 23.0	60.8 ± 26.0	75.9 ± 33.1	25.8 ± 11.5
PED-CED-alt [4]	99.2 ± 0.6	50.8 ± 23.6	62.5 ± 25.9	78.5 ± 32.2	24.6 ± 11.8
PED-CED [4]	99.4 ± 0.6	55.7 ± 25.0	67.7 ± 25.7	77.7 ± 32.8	22.2 ± 12.5
U -Net (ours)	**99.70 ± 0.33**	**77.10 ± 19.20**	**85.05 ± 16.05**	**88.37 ± 21.55**	**5.03 ± 2.24**

Our model delivers consistent gains over all CED-based baselines. Relative to the strongest prior (PED–CED), we observe higher Accuracy (99.70% vs. 99.4%), substantially better IoU (77.10% vs. 55.7%), higher Precision (85.05% vs. 67.7%), and improved Recall (88.37% vs. 77.7%). The E_2 error decreases from 22.2% to 5.03%, indicating far fewer false positives and false negatives overall. Standard deviations also decrease for key measures (e.g., IoU: 19.2% vs. 25.0%), suggesting improved stability across images. While gains over the original U -Net baseline are modest (IoU: 77.1% vs.

 doi:10.20944/preprints202604.1693.v1

76.2%), the proposed ResNet-50–based U-Net offers enhanced robustness to occlusion and illumination, stronger cross-dataset generalization (IoU: 94.76% on EarSegDB-25), and improved metric stability—highlighting architectural advantages beyond marginal improvements. Together, these results show that the proposed U-Net architecture yields markedly tighter overlap with ground truth, higher detection fidelity, and substantially lower error rates for pixel-wise ear segmentation in unconstrained conditions.

Figure 11 compares our U-Net (ResNet-50) with prior encoder–decoder baselines from [4]. All curves are computed at the pixel level by sweeping the mask threshold; the x-axis is restricted to 40–100% recall to match the plotting convention in [4]. Across almost the entire operating region (roughly 50–95% recall), our curve remains consistently above the baselines, indicating fewer false positives at comparable recall. In particular, our method sustains very high precision through the mid- and high-recall regime and only drops sharply near full recall, whereas SegNet and the PED–CED variants degrade steadily as recall increases. This dominance of the purple curve implies a larger area under the PR curve (higher AP) and reflects better calibration and separability of ear vs. background pixels in unconstrained images.

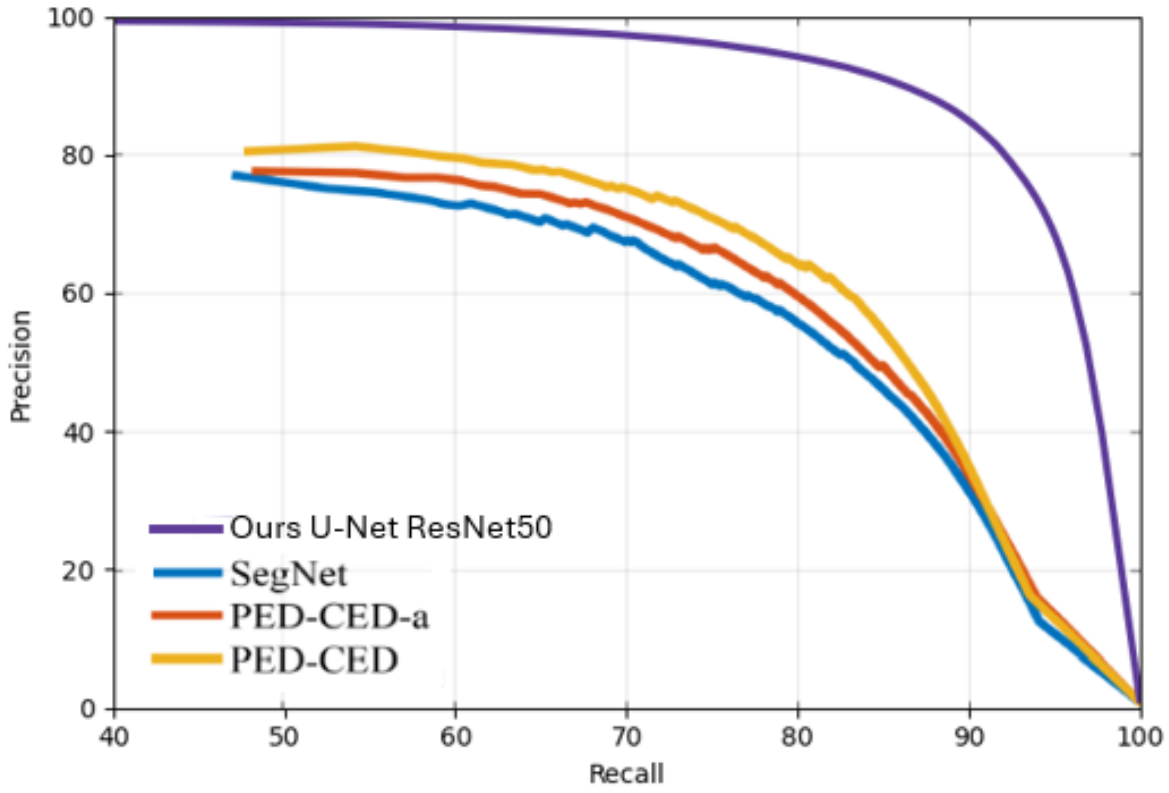

Figure 11. Precision–Recall curves on AWE (pixel level, 40–100% recall). Our U-Net (ResNet-50, purple) compared to SegNet and PED–CED baselines from [4]. Our curve stays higher over most of the recall range, indicating superior precision at matched recall and a larger area under the curve (AP).

Runtime comparison:Table 7 reports the average per-image processing time (lower is better) on the full test set for 480×360 px inputs. Our U-Net ResNet50 achieves the fastest runtime at **77.5 ms** per image—**2.3× faster** than the Haar baseline (178 ms), **8.8% faster** than SegNet (85 ms), and **12.9% faster** than both PED–CED variants (89 ms). At 77.5 ms, the method runs at roughly **12.9 fps**, offering a strong accuracy–latency trade-off.

 doi:10.20944/preprints202604.1693.v1

Table 7. Average per-image detection time (ms) on 480×360 px images. Lower is better. Times are averaged over the entire test set.

Approach	Detection Time (ms)
Haar[4]	178
SegNet[4]	85
PED–CED-alt[4]	89
PED–CED[4]	89
U-Net ResNet50 (ours)	**77.5**

Qualitative Results

Figure 12 shows representative predictions on the AWE test set under unconstrained conditions. In each panel, the red overlay is the predicted ear mask and the green contour is the ground-truth boundary. The per-image IoU is annotated. The model generally aligns well with the annotated pinna (e.g., IoU ≈ 86–90%), while lower IoUs (e.g., ∼59%) arise in challenging cases with severe pose, hair occlusion, or weak edge contrast. These examples complement the quantitative metrics by illustrating good boundary adherence across diverse subjects and scenes, together with typical failure modes in the wild. Faces are pixelated for display to preserve privacy.

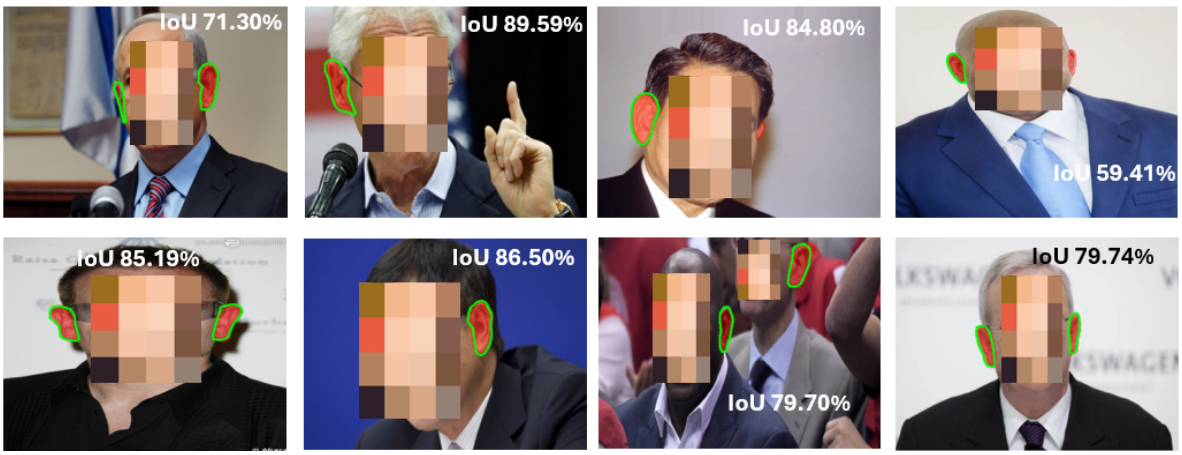

Figure 12. AWE test-set qualitative segmentation. Predicted ear masks (red) overlaid with ground-truth contours (green); each panel reports the per-image IoU. High IoUs demonstrate accurate boundary recovery across varied poses and lighting, while lower IoUs reflect difficult cases with occlusion and low contrast. Faces are anonymized via pixelation.

Figure 13 illustrates challenging AWE test images where the model underperforms (IoU ≈ 0–41%). The predicted mask (red) often drifts from the ground truth (green) due to strong out-of-plane pose, partial visibility of the ear, occlusions from hair or accessories, low resolution, motion blur, weak edge contrast, or distracting backgrounds. These examples highlight remaining sensitivities to boundary ambiguity, scale, and occlusion in the wild. In future work, we aim to mitigate such errors via occlusion/illumination/blur augmentations, higher-resolution and multi-scale features, boundary-aware losses, and light post-processing to refine edges.

 doi:10.20944/preprints202604.1693.v1

Figure 13. AWE failure cases with low overlap. Predicted masks (red) and ground-truth contours (green); per-image IoU shown in each panel. Typical errors arise from severe pose, occlusion, low contrast or blur, and distracting backgrounds. Faces are pixelated for privacy.

5. Discussion

Our study assessed a U-Net with a ResNet-50 encoder for automatic ear segmentation across both a controlled ear dataset (EarSegDB) and an unconstrained benchmark (AWE). Quantitatively, the method shows consistent improvements over prior convolutional encoder–decoder (CED) baselines from [4]. In the CED comparison (Table 6), our model achieves strong performance across all reported metrics, including **99.70±0.33%** accuracy, **77.10±19.20%** IoU, **85.05±16.05%** precision, **88.37±21.55%** recall, and a lower E_2 error of **5.03±2.24%**. On EarSegDB, we further observe high overlap scores (e.g., test IoU **94.76%**), with validation and training sets showing similarly strong performance. Notably, the test IoU being slightly higher than training/validation is consistent with reports in segmentation literature (e.g., U-Net [41], UNet++ [43]), and likely reflects distributional alignment and reduced variability in the test split rather than overfitting.

 doi:10.20944/preprints202604.1693.v1

As discussed in Section 3.10 and illustrated in Figure 8, pixel-wise Accuracy can be inflated in background-dominated ear images and therefore may overestimate true segmentation quality. Accordingly, we interpret performance primarily through overlap-based metrics such as IoU, which more faithfully capture the extent to which the predicted mask aligns with the ear region.

Precision–Recall analysis further supports these observations. When plotted following the convention of [4] (recall range 40–100%), our PR curve (Figure 11) remains consistently above SegNet and the PED–CED variants across most of the operating range (approximately 50–95% recall), indicating fewer false positives at comparable recall levels and a larger area under the curve (higher AP). For completeness, we report an *AP@40–100* of **54.8%**, which reflects the truncated recall range used in the figure; the underlying curve maintains high precision through the mid–high recall regime and decreases only near full recall.

Qualitative results corroborate the quantitative trends. On EarSegDB and AWE, the predicted masks adhere closely to the pinna boundary across diverse poses, subjects, and lighting (Figures 9 and 12). The failure analyses (Figures 10 and 13) reveal typical in-the-wild challenges: strong out-of-plane pose, partial visibility, hair and accessories occlusions, low contrast, motion blur, and distracting background textures. These factors widen the IoU distribution on AWE (standard deviation ≈19%), but do not change the overall comparative trend.

Runtime measurements (Table 7) indicate that the proposed approach is computationally efficient, achieving **77.5 ms** per 480×360 image on the test set—faster than Haar (178 ms), SegNet (85 ms), and both PED–CED variants (89 ms)—corresponding to approximately 12.9 fps and providing a favorable balance between accuracy and latency.

Limitations and avenues for improvement. Performance degrades under heavy occlusion, extreme pose, or weak edges, and the model can occasionally over- or under-segment boundary details. Dataset size and annotation noise may further limit peak IoU on challenging images. Future work should explore (i) targeted augmentations for occlusion/illumination/blur, (ii) boundary-aware and class-imbalance losses (e.g., Dice+BCE, focal/Tversky, or boundary loss), (iii) multi-scale and attention modules for improved context aggregation, (iv) lightweight post-processing (e.g., CRF or morphology) to refine edges, and (v) domain adaptation or semi/self-supervised learning to leverage unlabeled in-the-wild images. For deployment, model compression (quantization/pruning) and hardware-specific inference backends can further reduce latency.

6. Conclusion

We presented a U-Net with a ResNet-50 encoder for pixel-wise ear segmentation in both controlled and unconstrained settings. The approach demonstrates consistent improvements over CED baselines across multiple evaluation metrics, including accuracy, IoU, precision, recall, and E_2 error. Precision–Recall analysis indicates improved precision across much of the high-recall regime, and the method achieves an average processing time of **77.5 ms** per 480×360 image. Qualitative examples illustrate accurate boundary adherence across diverse subjects and lighting conditions, while failure cases highlight the remaining challenges of occlusion and low contrast in unconstrained environments. Overall, the results suggest that incorporating a residual encoder within a U-Net framework can improve segmentation performance in ear biometrics. While the architectural modification is incremental, the empirical results indicate its effectiveness as a practical solution for ear segmentation in biometric pre-processing pipelines.The primary contribution of this work lies in systematic evaluation across multiple datasets and metrics, providing insights into segmentation behavior under unconstrained conditions. Future work will focus on improving robustness under occlusion and pose variations, enhancing boundary precision, and further optimizing runtime performance for real-time and on-device applications.

Institutional Review Board Statement: This study uses publicly available datasets (AWE and EarSegDB-25), and no new human subject data were collected. Ethical approval and participant consent were handled by the original dataset creators.

 doi:10.20944/preprints202604.1693.v1

References

1. Emeršič, Ž.; Štruc, V.; Peer, P. Ear recognition: More than a survey. *Neurocomputing* **2017**, *255*, 26–39.
2. Ma, Y.; Huang, Z.; Wang, X.; Huang, K. An overview of multimodal biometrics using the face and ear. *Mathematical Problems in Engineering* **2020**, *2020*, 6802905.
3. Borbora, K.A.; Lahkar, R. Automatic segmentation of human ear in the wild. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)* **2024**, *34*, 333–341.
4. Emeršič, Ž.; Gabriel, L.L.; Štruc, V.; Peer, P. Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. *IET Biometrics* **2018**, *7*, 175–184.
5. Emeršič, Ž.; Štepec, D.; Štruc, V.; Peer, P.; George, A.; Ahmad, A.; Omar, E.; Boult, T.E.; Safdaii, R.; Zhou, Y.; et al. The unconstrained ear recognition challenge. In Proceedings of the 2017 IEEE international joint conference on biometrics (IJCB). IEEE, 2017, pp. 715–724.
6. Hossain, A.; Sultan, T.; Chowdhury, M.Z.U.; Schuckers, S. Deep Learning Approach for Ear Recognition and Longitudinal Evaluation in Children. In Proceedings of the 2024 International Conference of the Biometrics Special Interest Group (BIOSIG), 2024, pp. 1–7. https://doi.org/10.1109/BIOSIG61931.2024.10786753.
7. Karimi, D.; Dou, H.; Warfield, S.K.; Gholipour, A. Transfer learning in medical image segmentation: New insights, better results? *NeuroImage* **2021**, *232*, 117889.
8. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Eskeland, S.L. ResUNet++: An advanced architecture for medical image segmentation. In Proceedings of the IEEE International Symposium on Multimedia (ISM). IEEE, 2019, pp. 225–2255.
9. Alsabhan, A.; Alajlan, N.; El-Saban, M. Detecting buildings and nonbuildings from satellite imagery using U-Net with ResNet/VGG encoders. *Remote Sensing* **2022**, *14*, 1292.
10. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017, pp. 1–4.
11. Iglovikov, V.; Shvets, A. TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation. *arXiv preprint arXiv:1801.05746* **2018**.
12. University of Science and Technology Beijing (USTB). USTB Ear Database. http://www1.ustb.edu.cn/resb/en/index.htm.
13. University of Notre Dame. UND Computer Vision Research Lab (CVRL) Datasets. https://cvrl.nd.edu/projects/data/. Accessed: August 20, 2025.
14. Kumar, A.; Wu, C. Automated human identification using ear imaging. *Pattern Recognition* **2012**, *45*, 956–968.
15. Arbab-Zavar, B.; Nixon, M.S. On shape-mediated enrolment in ear biometrics. In Proceedings of the International Symposium on Visual Computing. Springer, 2007, pp. 549–558.
16. Ramos-Cooper, S.; Gomez-Nieto, E.; Camara-Chavez, G. Vggface-ear: an extended dataset for unconstrained ear recognition. *Sensors* **2022**, *22*, 1752.
17. EarVN1.0 Dataset. https://data.mendeley.com/datasets/yws3v3mwx3/4. Mendeley Data, Version 4. Accessed: August 20, 2025.
18. Zhang, Y.; Mu, Z.; Yuan, L.; Yu, C.; Liu, Q. USTB-Helloear: A large database of ear images photographed under uncontrolled conditions. In Proceedings of the International Conference on Image and Graphics. Springer, 2017, pp. 405–416.
19. Raposo, R.; Hoyle, E.; Peixinho, A.; Proença, H. UBEAR: A dataset of ear images captured on-the-move in uncontrolled conditions. In Proceedings of the 2011 IEEE workshop on computational intelligence in biometrics and identity management (CIBIM). IEEE, 2011, pp. 84–90.
20. Computer and Technology in Medicine Group (CTIM). AMI Ear Database. https://ctim.ulpgc.es/research_works/ami_ear_database. Accessed: August 18, 2025.
21. UERC: Unconstrained Ear Recognition Challenge. http://uerc.fri.uni-lj.si. Faculty of Computer and Information Science (FRI), University of Ljubljana. Accessed: August 20, 2025.
22. Frejlichowski, D.; Tyszkiewicz, N. The west pomeranian university of technology ear database–a tool for testing biometric algorithms. In Proceedings of the International Conference Image Analysis and Recognition. Springer, 2010, pp. 227–234.
23. Adebayo, A.A.; Elizabeth, B.; Olusegun, F.; Abayomi, D.G.; Bamidele, A.J. An Occlusion and Pose Sensitive Image Dataset for Black Ear Recognition, 2023. Accessed: Aug. 29, 2025, https://doi.org/10.5281/zenodo.7715969.
24. Lahkar, R.; Borbora, K.A. EarSegDB 25, 2023. Version 1, https://doi.org/10.17632/zp5c895yrg.1.

 doi:10.20944/preprints202604.1693.v1

25. Yan, P.; Bowyer, K. Empirical evaluation of advanced ear biometrics. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops. IEEE, 2005, pp. 41–41.
26. Alvarez, L.; González, E.; Mazorra, L. Fitting ear contour using an ovoid model. In Proceedings of the Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology. IEEE, 2005, pp. 145–148.
27. Deepak, R.; Nayak, A.V.; Manikantan, K. Ear detection using active contour model. In Proceedings of the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS). IEEE, 2016, pp. 1–7.
28. Ansari, S.; Gupta, P. Localization of ear using outer helix curve of the ear. In Proceedings of the 2007 International Conference on Computing: Theory and Applications (ICCTA'07). IEEE, 2007, pp. 688–692.
29. Prakash, S.; Gupta, P. An efficient ear localization technique. *Image and Vision Computing* **2012**, *30*, 38–50.
30. Prakash, S.; Gupta, P. Ear Detection in 2D. In *Ear Biometrics in 2D and 3D: Localization and Recognition*; Springer Singapore, 2015; pp. 21–49. https://doi.org/10.1007/978-981-287-739-0_2.
31. Chen, H.; Bhanu, B. Human ear detection from side face range images. In Proceedings of the Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, 2004, Vol. 3, pp. 574–577.
32. Said, E.H.; Abaza, A.; Ammar, H. Ear segmentation in color facial images using mathematical morphology. In Proceedings of the 2008 Biometrics Symposium. IEEE, 2008, pp. 29–34.
33. Cummings, A.H.; Nixon, M.S.; Carter, J.N. A novel ray analogy for enrolment of ear biometrics. In Proceedings of the 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2010, pp. 1–6.
34. Ibrahim, M.I.; Nixon, M.S.; Mahmoodi, S. Shaped wavelets for curvilinear structures for ear biometrics. In Proceedings of the International Symposium on Visual Computing. Springer, 2010, pp. 499–508.
35. Wahab, N.K.A.; Hemayed, E.E.; Fayek, M.B. HEARD: An automatic human EAR detection technique. In Proceedings of the 2012 international conference on engineering and technology (ICET). IEEE, 2012, pp. 1–7.
36. Yuan, L.; Lu, F. Real-time ear detection based on embedded systems. In Proceedings of the 2018 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2018, Vol. 1, pp. 115–120.
37. Zhang, Y.; Mu, Z. Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. *Symmetry* **2017**, *9*, 53.
38. Zhang, Y.; Mu, Z.; Yuan, L.; Zeng, H.; Chen, L. 3D ear normalization and recognition based on local surface variation. *Applied Sciences* **2017**, *7*, 104.
39. Ribič, M.; Emeršič, Ž.; Štruc, V.; et al. Influence of alignment on ear recognition: case study on AWE dataset. In Proceedings of the International electrotechnical and computer science conference, 2016, Vol. 25, pp. 131–134.
40. Emeršic, Ž.; Peer, P.; Dimitrovski, I. Assessment of predictive clustering trees on 2D-image-based Ear recognition. In Proceedings of the International Electrotechnical and Computer Science Conference, 2016.
41. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
43. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the International workshop on deep learning in medical image analysis. Springer, 2018, pp. 3–11.