Image and Vision Computing 154 (2025) 105424

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Synthesizing multilevel abstraction ear sketches for enhanced biometric recognition

David Freire-Obregón [a,*], Joao Neves [b], Žiga Emeršič [c], Blaž Meden [c], Modesto Castrillón-Santana [a], Hugo Proença [b]

[a] Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
[b] Universidade da Beira Interior, Covilha, Portugal
[c] University of Ljubljana, Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Ear biometrics
Sketch-based identification
Triplet-loss function
Cross-dataset generalizability

ABSTRACT

Sketch understanding poses unique challenges for general-purpose vision algorithms due to the sparse and semantically ambiguous nature of sketches. This paper introduces a novel approach to biometric recognition that leverages sketch-based representations of ears, a largely unexplored but promising area in biometric research. Specifically, we address the "*sketch-2-image*" matching problem by synthesizing ear sketches at multiple abstraction levels, achieved through a triplet-loss function adapted to integrate these levels. The abstraction level is determined by the number of strokes used, with fewer strokes reflecting higher abstraction. Our methodology combines sketch representations across abstraction levels to improve robustness and generalizability in matching. Extensive evaluations were conducted on four ear datasets (AMI, AWE, IITDII, and BIPLab) using various pre-trained neural network backbones, showing consistently superior performance over state-of-the-art methods. These results highlight the potential of ear sketch-based recognition, with cross-dataset tests confirming its adaptability to real-world conditions and suggesting applicability beyond ear biometrics.

1. Introduction

Facial recognition technology has become an invaluable asset for law enforcement agencies, offering significant assistance in forensic investigations [1]. Nevertheless, obtaining direct photographs of suspects might not always be possible in forensic settings. In such cases, the assistance of eyewitnesses or victims becomes invaluable, as they can help in generating a facial sketch. Although these sketches provide an approximate representation of the suspect's identity, they have a crucial role in narrowing down the list of potential suspects by facilitating searches within the mugshot database.

Facial sketches are mainly divided into two distinct categories [2]: composite and hand-drawn. Composite sketches are produced using specialized software, and offer a streamlined and technologically driven approach. Conversely, hand-drawn sketches yield from proficient forensic artists, which requires significant training and expertise. This kind of methods relies on the artist's ability to translate eyewitness descriptions into accurate visual representations.

Contrary to popular belief, police sketch artists often create multiple sketches of suspects, not just from one point of view or in a single style [3–5]. The variety in sketch representations, including side-view sketches, is essential for comprehensively illustrating a suspect's features. While frontal sketches highlight the face's direct features, side views reveal crucial angular details like ear shape, jawline, and nose shape, offering a complete profile that enhances the likelihood of identification. Side-view sketches are invaluable in criminal investigations, especially in the age of ubiquitous surveillance. CCTV often captures profiles, not frontals, making these sketches align well with footage, thereby boosting law enforcement's ability to match suspects.

Significant efforts have been directed towards addressing the complexities of frontal face sketch recognition [6–9]. Although sketches might appear similar to photographs at first glance, discernible distinctions between the two remain prevalent. Such differences are mainly attributable to the challenges in accurately recalling and rendering a face from memory, a phenomenon acknowledged as the *modality gap*. This gap underscores a crucial obstacle that impacts the accuracy of sophisticated face recognition algorithms, especially in contexts requiring cross-modal comparisons between sketches and photographic images [10,11]. In response, some researchers advocate for the adoption of component-based strategies in facial sketch recognition like eyes, nose, mouth, and forehead [12–14]. However, these methodologies have predominantly focused on frontal facial sketches, neglecting

* Corresponding author.
E-mail address: david.freire@ulpgc.es (D. Freire-Obregón).

https://doi.org/10.1016/j.imavis.2025.105424
Received 21 September 2024; Received in revised form 1 December 2024; Accepted 12 January 2025
Available online 18 January 2025

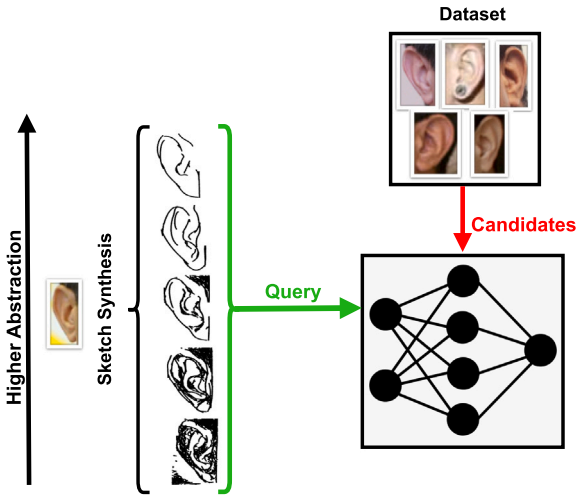


Fig. 1. Proposed multi-abstraction sketch-ear recognition scheme. Left: generating sketches across various abstraction levels involves creating representations with varying levels of detail, encompassing stroke counts from 16, 32, 64, 128, down to 256 shown in the figure from top to bottom. It is important to note that strokes are not cumulative; higher-level syntheses do not include strokes repeated in lower-level syntheses. Right: the original photos (the sketched photo is removed from the set). Bottom right: the proposed pipeline to efficiently recognize a subject in a single forward pass.

the potential of ears. In this work, we focus on exploring ear sketches, motivated by the ear's intricate and individual-specific structure, which remains consistent over time, offering a distinctive and reliable marker for suspect identification [15].

We meticulously processed four distinguished ear datasets (AMI, AWE, IITDII, and BIPLab), each characterized by its unique acquisition procedure, thus expanding the scope of our research. We transformed each dataset into multiple abstracted versions using different numbers of strokes (16, 32, 64, 128, and 256) for sketching (see Fig. 1). Utilizing these abstracted versions alongside the original RGB data, we assessed various pre-trained models to enhance a multi-input triplet-loss function, with each input fine-tuning the embedding generator at distinct abstraction levels. This method stands out for its ability to leverage insights from each abstraction level, enhancing the accuracy and effectiveness of recognition tasks. Our contributions can be outlined as follows:

- We have implemented the approach to analyze lateral view components of the face, such as ears, which have received limited attention in the literature. This aspect is particularly significant because lateral views provide distinctive biometric features that can improve recognition systems' overall accuracy and robustness.
- We have expanded the ear datasets mentioned by generating their sketched versions with CLIPasso [16], which converts an image of an object to a sketch, allowing for varying levels of abstraction while preserving its key visual features.
- We adapted the triplet-loss function to integrate multiple abstraction levels, showcasing its enhanced performance over traditional baselines and when abstraction levels are considered in isolation. This is significant as it leverages the composite strengths of various abstract representations, yielding more refined and accurate recognition capabilities.
- In the scope of this study, we perform an exhaustive comparative analysis of various backbones in our proposed pipeline, evaluating their efficacy in sketch-based ear recognition.
- We conducted cross-dataset experiments to evaluate the generalizability of our approach, training on a combination of datasets while testing on a distinct one. This setup simulates real-world scenarios where models encounter unseen data distributions, revealing the robustness and adaptability of our method compared to baselines.

2. Related work

Sketch Synthesis. Facial sketch recognition approaches can be broadly categorized into generative and discriminative approaches [8]. Generative approaches learn a modality transfer function to ensure matching is carried out in the same modality [17,18]. In contrast, discriminative approaches focus on feature extraction such as scale-invariant feature transform (SIFT) [19] or multiscale local binary pattern (MLBP) [20]. However, the adequacy of these features for cross-modal recognition tasks often needs to be improved [21], leading to a pivot towards techniques that ascertain or learn features invariant across modalities [7,22].

Significant advances have been made in face sketch recognition in recent years, with early methods leveraging deep learning for enhanced accuracy. Mittal et al. [23] introduced a transfer learning strategy by training a deep network trained on a large number of photos, and subsequently training it with a reduced amount of sketch-photo pairs bridging the gap created by the scarcity of extensive sketch-photo datasets. On another research line, pre-trained face recognition models were used to identify sketches, by extracting features from intermediate layers and adopting a simple metric learning approach rather than comprehensive model fine-tuning, yielding results comparable or superior to previous methods [11].

Ear Recognition. Considering the existing body of work in facial sketch recognition, it is intriguing to consider the unexplored potential of applying these methodologies to ear sketches. To the best of our knowledge, ear sketches have not been addressed explicitly in the literature, representing a novel avenue for research. Ears, with their unique and complex structures, offer a distinctive biometric feature that remains consistent over time, much like fingerprints. This consistency and the distinctiveness of ear shapes and features make ear sketches an interesting and potentially fruitful focus for sketch recognition efforts. In this regard, ears have shown a remarkable recognition performance over the past two decades. Since the pioneering work in which facial and ear eigenvectors were combined to boost performance over individual biometrics [24], several authors have considered the ear as a biometric trait. Over time, numerous studies have addressed ear recognition, even under unconstrained conditions [25]. Initially, the focus was on 2D ear images, and a survey categorized recognition techniques into geometric, holistic, local, and hybrid methods [26]. Geometric methods leverage the ear's geometric features [27], while holistic strategies view the ear entirely, extracting global property features, like the force fields method [28]. Local methods focus on extracting features from specific image regions for recognition [29], and hybrid methods blend elements from the approaches mentioned above [30]. Similar to the advancements in facial sketch recognition, recent studies have showcased the transformative capabilities of deep learning [31,32], outperforming traditional methods reliant on manually engineered features [30,33].

In this work, we aim to bridge the gap in the existing literature by focusing on ear sketches and their recognition on real images, an area that has yet to be extensively explored. While the methodologies for facial sketch recognition have matured, the unique contours and features of the ear present different challenges and opportunities for biometric recognition. Deep learning approaches, which have revolutionized facial sketch recognition, also hold significant promise for ear sketch recognition. By leveraging deep learning techniques and the rich yet underexplored biometric information present in ear sketches, we strive to develop robust recognition systems that accurately identify individuals based on their ear sketches.

3. Methodology

3.1. Overview

Let I be an image of an ear in a high-dimensional space $\mathbb{R}^n$. We consider a sketch $S(I)$ of the ear as a transformation of I into a lower-dimensional space $\mathbb{R}^m$ $(m < n)$, which captures the essential features of

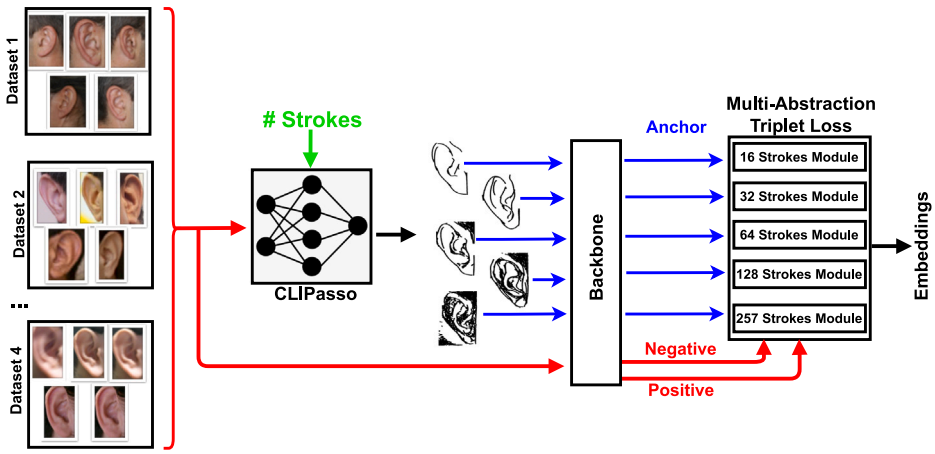


Fig. 2. Overview of the experimental workflow. The process begins with four datasets (AMI, AWE, IITDII, BIPLab) transformed into sketches with varying abstraction levels using CLIPasso (16 to 256 strokes). The triplet-loss framework processes anchor sketches, positive, and negative samples to generate embeddings.

I while also considering the sketch as a set of strokes:

$$S : \mathbb{R}^n \to \mathbb{R}^m, \quad S(I) = \bigcup_{i=1}^{N} s_i = g_1(I), g_2(I), \ldots, g_m(I), \tag{1}$$

where $S(I)$ is the sketch representation of the ear image, s_i represents the ith stroke contributing to the sketch, and N is the total number of strokes. Note that this latter parameter is inversely related to the abstraction level A of the sketch, indicating that a higher number of strokes (or features g_j) corresponds to a lower level of abstraction and vice versa. Moreover, the features $g_j(I)$ of each stroke s_i are determined based on visually salient and semantically meaningful components of the image [16]. Each stroke's parameters are optimized through an iterative process that balances geometric coherence with semantic fidelity, using control points to represent essential contours and shapes within a lower-dimensional space. This approach ensures that the sketch retains key visual characteristics of the original image while achieving the desired level of abstraction. The concept of stroke-related abstraction has recently been discussed in the literature, particularly in the image retrieval domain [34].

Then, as shown in Fig. 2, we redefine the sketch representation $S(I)$ of an ear image I as the embeddings produced by pre-trained neural networks, deviating from the traditional notion of sketches as sets of strokes. We employ eight state-of-the-art backbones, namely VGG19, ResNet152, Xception, InceptionV3, MobileNet, EfficientNetB7, EfficientNetB0, and DenseNet121, each trained on the ImageNet dataset. These networks, known for their distinct architectural features—from VGG19's deep convolutional layers to ResNet152's skip connections, and from the modular design of InceptionV3 and Xception to the efficient and scalable architecture of EfficientNets—transform the high-dimensional input image $S(I)$ in $\mathbb{R}^m$ into a lower-dimensional, dense embedding in $\mathbb{R}^p$ ($m <= p$), capturing the essential features of the ear sketch $S(I)$ defined in Eq. (1).

$$B : \mathbb{R}^m \to \mathbb{R}^p, \quad B(I) = b(S(I)) \in \mathbb{R}^p. \tag{2}$$

Here, B represents the backbone network that acts as a function mapping the high-dimensional image to its embedding, providing a nuanced feature-rich representation. This interpretation of an ear sketch as an embedding enables us to leverage the advanced feature extraction capabilities developed through deep learning, offering a robust and discriminative representation of ear images suitable for complex biometric recognition tasks.

Finally, with a query ear sketch denoted as $b(S(I))$, and a collection of M candidate photos $b(c_j)_{j=1}^{M}$ within C, our goal is to evaluate the similarity between $b(S(I))$ and $b(c)$ to rank the entire gallery of photos. This prioritization aims to highlight the authentic match for the query sketch. The task poses two primary challenges: (i) closing the domain gap between sketches and photos, and (ii) accurately distinguishing subtle discrepancies among candidate photos to achieve precise ranking despite the domain gap and the inherent variability of sketches. To address these challenges, we propose employing a triplet network model to acquire a domain-invariant representation $f_\theta(\cdot)$, where θ represents a configuration of abstraction levels. These levels can be combinations such as 16, $16 \cup 32$, $16 \cup 32 \cup 64$, etc., where $\cup$ signifies the union of different abstraction levels considered in the architecture. This representation simplifies the measurement of similarity between $b(S(I))$ and $b(c) \in C$ using the Euclidean distance equation: $D(b(S(I)), b(C)) = \|f_\theta(b(S(I))) - f_\theta(b(C))\|_2^2$.

3.2. Sketch generation

Sketch generation was obtained using CLIPasso [16] for all the ear datasets considered. Each sketch was carefully designed to consist of 16, 32, 64, 128, or 256 pen strokes. Unlike other studies, we opted for a comparatively higher number of strokes, a decision influenced by the demands of the recognition task [35]. The process employed by CLIPasso involves adjusting the parameters of various curves, including start/end points and control points. Each curve represents a single pen stroke, and the adjustment is made to mimic the target image accurately. To guide this adjustment process, we relied on a pre-trained implementation of CLIP [16]. This model, developed through contrastive learning on an extensive dataset of text-image pairs, provided valuable insights into achieving similarity between the generated sketch and the target image. Similarity is determined based on the distance computed between CLIP's embedding of the target image and the embedding of the sketch. These embeddings capture a combination of feature activations from multiple intermediate layers of CLIP, offering a comprehensive representation of the visual and semantic characteristics of both the target image and the generated sketch.

3.3. Multi-abstraction triplet loss

Our network architecture introduces a novel approach by employing multiple modules within a single triplet framework, each module dedicated to processing a sketch at a different level of abstraction. As shown in Fig. 3, the framework comprises three branches: one for the query sketch, processed through several modules to handle varying abstraction levels, and two additional branches for the positive and negative photos, which remain consistent across all modules. This multi-module strategy allows for generating multiple embeddings, one from each module, which are then averaged to form a unified embedding used in the triplet configuration. While innovative in our context, this method

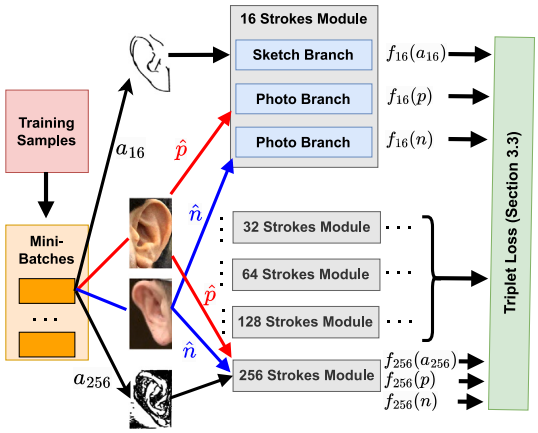


Fig. 3. Architecture of the proposed learning network. Our architecture integrates multiple modules within a single triplet framework, processing anchor sketches at varying levels of abstraction and ensuring consistency with positive and negative images. The anchor is a sketched image of the same subject as the positive image (but from a different photo), while the negative image represents a different subject. All images are processed at the abstraction level corresponding to the anchor's sketch, with the resulting embeddings fed into the customized triplet loss function.

builds on the concept of utilizing sketch-photo combinations for image retrieval, which has been previously explored [36,37]. However, unlike traditional applications in face recognition, which are typically confined to controlled datasets with high-quality sketches [38], our approach extends to less constrained scenarios, ensuring robustness by not using the photo that generated the sketch as a positive sample within any of the modules.

The framework shown in Fig. 3 is designed to ensure that the photo branches across all modules share weights, promoting consistency. In contrast, the alignment of weights between the sketch and photo branches depends on whether the module is Siamese or heterogeneous [37]. This distinction is essential given the significant differences between sketch and photo domains, which underscore a *machine modality gap*. While heterogeneous modules are typically favored for distinct domains like text and images [39], Siamese modules are preferred for more closely related domains [40]. Previous studies have exploited heterogeneous modules for sketch-photo tasks [41]. However, later findings suggest such configurations may not be ideal for detailed sketch-based image tasks, especially with sparse training data [37]. Consequently, to address the overfitting linked to sparse data, our methodology employs a distinct architecture where each embedding module maintains internal parameter consistency but operates independently without sharing parameters with other embedding modules.

In the context of the triplet loss for ear sketches, considering multiple levels of abstraction, we define the triplet loss as follows. For each anchor embedding a_i in $\{a_1, a_2, \ldots, a_M\}$, where i represents different abstraction levels considered at a given configuration θ, with positive embedding p and negative embedding n:

$$L_{\text{acc}} = \sum_{i=1}^{M} \max(\|f_i(a_i) - f_i(\hat{p})\|^2 - \|f_i(a_i) - f_i(\hat{n})\|^2 + \alpha, 0), \tag{3}$$

where: $f_i(a_i) = f_i(b(S(I)))$ represents the ith new embedding of the anchor sketch at an i abstraction level. $f_i(\hat{p}) = f_i(b(p))$ is the ith new embedding of the positive photo embedding, similar to the anchor. $f_i(\hat{n}) = f_i(b(n))$ represents the ith new embedding of the negative photo embedding, dissimilar to the anchor and α is the margin enforced between positive and negative pairs.

This formulation enhances the recognition system's robustness by leveraging the intricate and unique features of ear sketches, considering multiple perspectives or variations of the anchor derived from the levels of abstraction in the sketch representation. Moreover, the architecture of each branch to generate $f_i(x)$ (see Fig. 3) consists of a neural module. Each input vector (a_i, $\hat{p}$, and $\hat{n}$) is fed into a dense layer with half the number of units as the input length, utilizing Rectified Linear Unit (ReLU) activation function, L2 kernel regularization with a coefficient of 1×10^{-3}, and He uniform initializer. Following the dense layer, batch normalization is applied to normalize the activations, followed by ReLU activation to introduce non-linearity. Dropout regularization with a rate of 0.5 is applied to mitigate overfitting. Subsequently, another dense layer with the same number of units as the previous one is employed, with similar configurations of activation, regularization, and initialization. After batch normalization, a third dense layer with the same number of units is utilized, employing identical configurations. Finally, the output layer consists of a dense layer with 512 units. This layer does not have an activation function, and the kernel weights are regularized using L2 regularization with a coefficient of 1×10^{-3} and initialized using the He uniform initializer. The output is then normalized using L2 normalization along the last axis.

4. Experimental setup

Datasets. Ear datasets exhibit substantial diversity in their sources, containing captured images and those obtained through web crawlers. These datasets also vary regarding ear pose (frontal or profile), participant count, dataset size, ethnicity representation, and camera settings. We have curated four datasets (AMI, IITDII, AWE, and BIPLab) to encompass various acquisition process variations. *AMI* [42] consists of 100 subjects, each with seven noiseless images. These images were captured under fixed illumination conditions using a 135 mm and 200 mm focal length. While yaw poses show minimal variation, pitch ranges notably from 40–45°. *IITD-II* [43] is a collection of grayscale images featuring 221 subjects captured under indoor lighting conditions with a fixed camera position to maintain consistent profile angles. All ear images are cropped, centered, and aligned, with images per user ranging from 3 to 6 samples. The *AWE* dataset [26] pioneers the concept of ear images captured in the wild by compiling images of 100 celebrities sourced from the internet under diverse conditions. Each subject contributes ten images, with sizes ranging significantly from 15×29 pixels to 473×1022 pixels. More recently, *BIPLab* [33] introduced a dataset comprising 300 images from 100 distinct participants. Diverging from traditional collections, images were captured under uncontrolled lighting and with a movable camera position. The dataset aims to simulate the ear portion captured during phone calls, covering approximately 90% of the image. Samples may exhibit blurring, with minimal yaw and pitch variation in ear poses.

Data Augmentation. Some analyzed datasets provide only a small number of samples per subject. For instance, the BIPLab dataset usually contains just three images per subject. To overcome this limitation, a data augmentation approach was implemented to expand the dataset, resulting in a threefold increase in samples per subject. This augmentation process involves diverse transformations, including random adjustments in brightness and contrast, horizontal flipping, shifting, scaling, and rotation [44]. It is worth noting that these augmented subsets are exclusively used for training purposes.

Metrics. We have employed two evaluation metrics for our analysis. Mean Average Precision (mAP) quantifies the average precision across all potential rankings of the images. Specifically, mAP is computed by determining each class's average precision (AP) and subsequently averaging these AP values across all classes. AP is the area under the precision–recall curve (PR curve) corresponding to a given query image. Additionally, we utilized the Cumulative Matching Characteristic (CMC) curve. This metric assesses the percentage of correct matches at each retrieved image rank. The CMC curve is constructed by computing the percentage of correct matches for each rank and plotting these results on a graph.

Implementation Details. In all our experiments, the initial learning rate was set to 0.001, with a mini-batch size of 128. The margin parameter α was set to 0.2.

Table 1
Mean average precision achieved by each backbone for the optimal abstraction level θ across different datasets: AMI and BIPLab with 64∪128∪256.

	Encoding	Dataset: AMI				Dataset: BIPLab			
		mAP↑	Rank-1↑	Rank-5↑	Rank-10↑	mAP↑	Rank-1↑	Rank-5↑	Rank-10↑
	LBP_{base}	16.2%	4.3%	24.0%	45.6%	15.8%	5.3%	23.3%	42.0%
	$Backbone_{base}$	17.8%	6.3%	24.7%	46.0%	18.0%	6.3%	24.6%	47.6%
Ours	VGG19	31.2%	16.1%	46.2%	69.1%	29.0%	15.0%	44.1%	69.0%
	ResNet152	35.9%	20.1%	51.2%	75.0%	32.2%	16.0%	45.1%	69.0%
	InceptionV3	31.7%	17.0%	48.1%	68.1%	29.7%	16.0%	38.2%	61.0%
	Xception	32.3%	18.0%	45.0%	67.1%	27.3%	14.0%	42.0%	59.1%
	DenseNet121	**40.6%**	**24.2%**	**64.7%**	**81.3%**	**39.2%**	**23.0%**	**60.3%**	**78.1%**
	EfficientNetB0	39.2%	23.0%	55.0%	78.0%	30.8%	14.5%	42.0%	69.0%
	EfficientNetB7	35.3%	20.4%	53.3%	73.1%	37.4%	20.0%	54.4%	71.1%
	MobileNet	36.5%	23.3%	48.2%	67.0%	30.7%	14.1%	45.2%	75.1%

Table 2
Mean average precision achieved by each backbone for the optimal abstraction level θ across different datasets: AWE and IITDII with 16∪32 ∪ 64 ∪ 128∪256.

	Encoding	Dataset: AWE				Dataset: IITDII			
		mAP↑	Rank-1↑	Rank-5↑	Rank-10↑	mAP↑	Rank-1↑	Rank-5↑	Rank-10↑
	LBP_{base}	16.2%	4.3%	24.0%	45.6%	15.8%	5.3%	23.3%	42.0%
	$Backbone_{base}$	17.8%	6.3%	24.7%	46.0%	18.0%	6.3%	24.6%	47.6%
Ours	VGG19	27.6%	13.0%	39.2%	66.1%	24.7%	10.5%	40.0%	60.9%
	ResNet152	27.8%	12.0%	40.1%	59.1%	26.2%	11.8%	41.4%	58.2%
	InceptionV3	24.9%	12.0%	35.1%	57.1%	23.7%	11.8%	33.2%	51.4%
	Xception	27.0%	14.0%	36.1%	58.0%	25.6%	12.7%	37.7%	56.8%
	DenseNet121	**33.8%**	**20.3%**	**46.2%**	**67.3%**	**33.8%**	**20.0%**	**51.4%**	**69.1%**
	EfficientNetB0	27.6%	11.1%	43.1%	66.0%	27.0%	14.5%	38.6%	54.1%
	EfficientNetB7	28.4%	16.1%	38.2%	58.1%	29.0%	15.0%	39.5%	61.8%
	MobileNet	29.2%	16.2%	38.1%	60.0%	26.2%	11.8%	41.4%	61.8%

5. Experimental evaluation

The results outlined in this section are based on the average accuracy obtained from five iterations of 4-fold cross-validation for each experiment. This methodology, commonly employed in previous studies [31,45–47], divides the dataset into four folds, each containing an equal number of samples. Typically, there are 25 subjects per fold, although the IITDII dataset deviates from this norm with 55 subjects per fold. This procedure is repeated five times. The performance metrics, including mAP and the CMC curve, are then averaged across all folds to obtain the final evaluation scores.

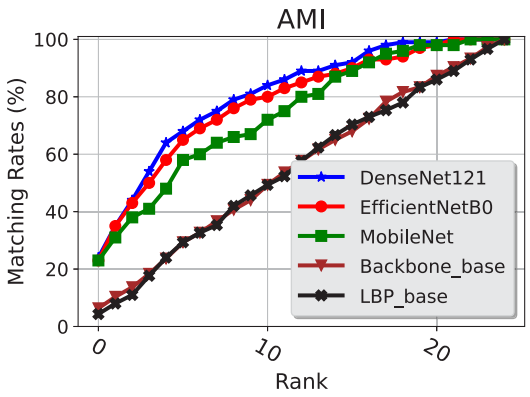
Different abstraction levels (θ as discussed in Section 3.1) were evaluated, including individual levels such as 16, 32, 64, 128, and 256 strokes, as well as combined configurations. These combined configurations encompass a diverse range of stroke counts, including 16∪32∪64, 16∪64∪256, 64∪128∪256, and 16∪32∪64∪128∪256, providing insights into the performance across different levels of detail in the sketches.

Also, we evaluate two baseline algorithms using implementations provided in the participants' starter kit for the Unconstrained Ear Recognition Challenge - UERC 2019. The first baseline algorithm, referred to as LBP_{base}, relies on a Local Binary Pattern (LBP)-based approach. Here, feature vectors for each test sample are computed directly from hand-crafted LBP features without any training [48,49]. These feature vectors are represented as histograms and compared using the Bhattacharyya distance. The second baseline algorithm, hereinafter denoted as $DenseNet121_{base}$, uses the DenseNet121 architecture for generating embeddings for test samples, which are then compared using the Euclidean distance. The choice of DenseNet121 as the backbone baseline is attributed to its superior performance across all datasets.

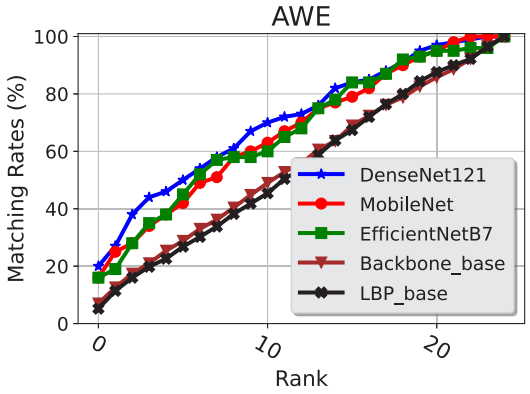
Tables 1 and 2 showcase the mAP attained by each backbone model at the optimal abstraction level θ across the considered datasets. Notably, DenseNet121 emerges as the frontrunner, yielding remarkable mAP scores of 40.6% for AMI, 33.8% for AWE, 33.8% for IITDII, and 39.2% for BIPLab. This consistent performance across datasets exhibits its prowess in extracting discriminative features from ear sketches. Despite DenseNet121's dominance, it is important to acknowledge the runner-up approaches, including EfficientNetB0, EfficientNetB7, and MobileNet, demonstrating competitive performance across both datasets. Additionally, it is noteworthy that DenseNet121 also excels in rank-1 accuracy, achieving commendable scores of 24.2% for AMI, 20.3% for AWE, 20.0% for IITDII, and 23.0% for BIPLab — solidifying its position as a reliable choice for accurate ear sketch recognition. Moreover, the consistent dominance of DenseNet121 underscores its robustness and adaptability to the challenges posed by real-world datasets, making it a compelling choice for practical deployment in diverse scenarios.

Fig. 4(a) shows the CMC curves for various models applied to the AMI dataset, revealing distinct model performance trends. DenseNet121 starts with an initial recognition rate of 24% and consistently improves until achieving an ideal recognition rate. This behavior underscores DenseNet121's adeptness in harnessing the dataset's uniformity to refine its identification accuracy progressively. Similarly, EfficientNetB0 embarks on a parallel path, achieving complete recognition with minor mid-rank fluctuations, indicating its resilience and effective learning from such a refined dataset. In contrast, MobileNet starts at 23% and exhibits a more gradual increase, suggesting a potential need for additional data to match the certainty levels of the other models in this dataset. Transitioning to Fig. 4(b), the narrative shifts to the AWE ear dataset, renowned for its challenging in-the-wild images. Here, DenseNet121 excels with a swift CMC curve escalation, affirming its capability to manage the dataset's inherent variability. MobileNet and EfficientNetB7 similarly demonstrate improvement, though their paths differ, emphasizing their distinct feature extraction and generalization capabilities. Meanwhile, the baseline models, $Backbone_{base}$ and LBP_{base}, display a more incremental learning curve, underscoring the nuanced challenges presented by these dataset's complexity.

Fig. 5(a) shows the CMC curves for various models on the IITDII dataset, which comprises grayscale images of Indian individuals and showcases the models' performance in recognizing a diverse set of features within a specific demographic. DenseNet121 starts at a 20% recognition rate and exhibits a consistent increase, reaching near-perfect identification, which underscores its effectiveness in adapting to the dataset's diversity and grayscale nature. EfficientNetB7 begins at 15%, showing a steady ascent in recognition capabilities, illustrating its adaptability and nuanced learning from the dataset's distinctive

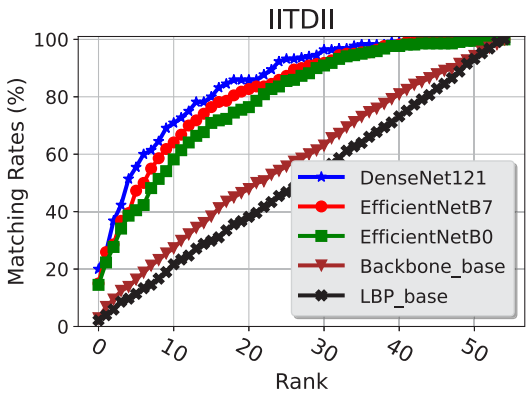


(a) CMC for AMI with baseline methods and top-3 mAP backbones.

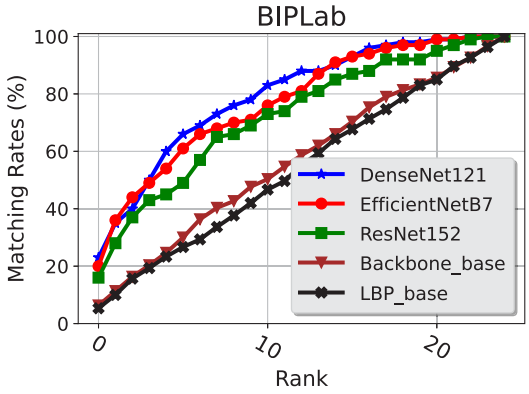


(b) CMC for AWE with baseline methods and top-3 mAP backbones.

Fig. 4. CMC curves for AMI and AWE, when compared to the performance of the top-performing models.



(a) CMC for IITDII with baseline methods and top-3 mAP backbones.



(b) CMC for BIPLab with baseline methods and top-3 mAP backbones.

Fig. 5. The CMC for IITDII and BIPLab showcase the performance of the top-performing models.

characteristics. EfficientNetB0, initiating at a slightly lower rate, progresses methodically, emphasizing its capability to extract relevant features from a dataset rich in cultural diversity. The $Backbone_{base}$ and LBP_{base} baselines, starting from much lower initial recognition rates, demonstrate a gradual improvement, reflecting again a more deliberate path to understanding the dataset's complexities.

In Fig. 5(b), the CMC curves for different models are presented against the backdrop of the BIPLab dataset, which is designed to mimic the ear region captured during phone calls with a focus on realism through blurring and minimal ear pose variations. DenseNet121 begins with a 23% identification rate, progressively climbing to perfect recognition, highlighting its robustness in handling images where the ear occupies a dominant portion and where blurring is prevalent. EfficientNetB7 shows a similar resilience, starting at 20% and methodically moving to full accuracy. Its performance trajectory emphasizes the model's capacity to navigate through the dataset's peculiarities, such as the limited yaw and pitch variations and the blurriness that characterizes the images. ResNet152, while starting at a lower initial recognition rate of 16%, demonstrates a steady improvement, reflecting its adaptability to the dataset's specific conditions, albeit at a slightly slower pace compared to DenseNet121 and EfficientNetB7. Again, the $Backbone_{base}$ and LBP_{base} baselines, with even lower starting points, exhibit a consistent rise in their CMC curves but a poorly performance.

6. Abstraction analysis

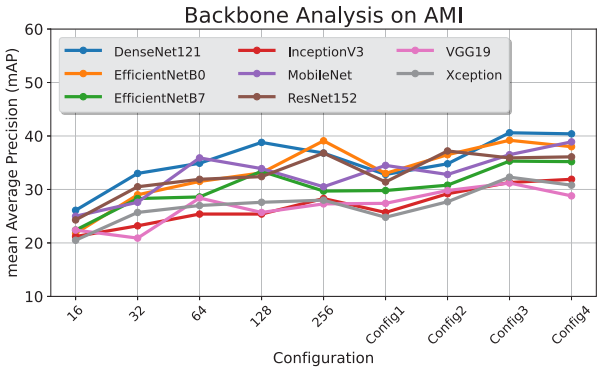
In this section, we focus on the impact of varying abstraction levels on the performance of our sketch-based recognition system. Specifically, we investigate how different levels of detail in the sketches—measured by the number of strokes—affect recognition accuracy across multiple datasets and backbone architectures. By analyzing the performance across different abstraction configurations, we aim to uncover insights into how the model handles varying levels of complexity in sketch representation and identify optimal configurations for improved recognition accuracy.

The analysis is divided into two parts. First, we evaluate how different backbone architectures respond to these varying levels of abstraction, providing a detailed comparison of their effectiveness under different conditions. Then, we examine the influence of abstraction levels across distinct datasets, each with unique characteristics and challenges.

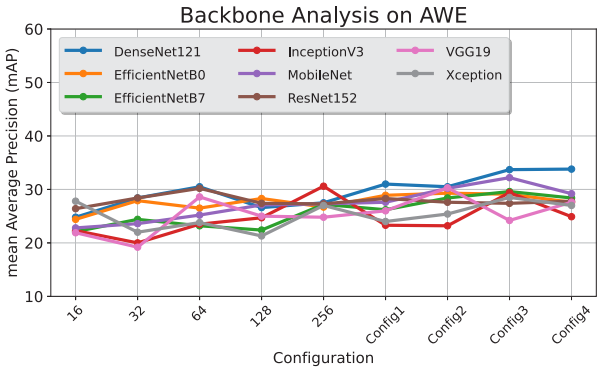
6.1. Backbones vs abstraction levels

Understanding how individual backbone architectures respond to varying levels of abstraction is a key aspect of our analysis. Each backbone has unique strengths and weaknesses when processing sketches at different levels of detail. Examining their performance across abstraction levels provides valuable insights into their suitability for the specific task of sketch-based ear recognition.

This section delves into the performance of various backbone models described in Section 3.1 at each abstraction level. We evaluate the effectiveness of each architecture across different stroke counts. The configurations, ranging from single stroke levels (16, 32, 64, 128, 256) to combined stroke levels (Config1: 16∪32∪64, Config2: 16∪64∪256, Config3: 64∪128∪256, and Config4: 16∪32∪64∪128∪256), illustrate how integrating multiple abstraction levels impacts model performance. By doing so, we aim to identify the most effective backbone for each abstraction level and highlight any performance trends that

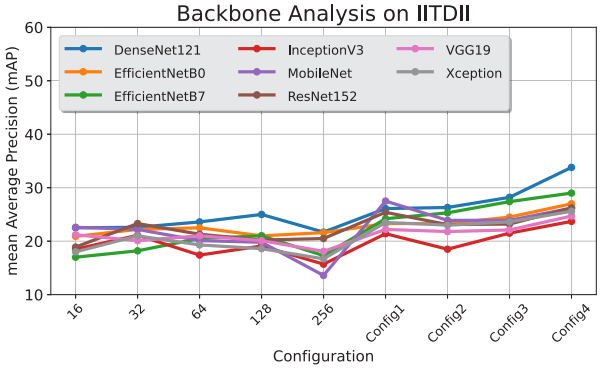


(a) AMI backbone mAP-analysis for each configuration.

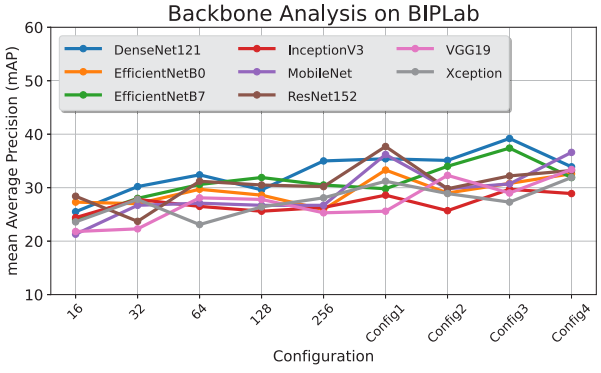


(b) AMI backbone mAP-analysis for each configuration.

Fig. 6. mAP analysis for AMI and AWE showcase the performance of each backbone.



(a) IITDII backbone mAP-analysis for each configuration.



(b) BIPLab backbone mAP-analysis for each configuration.

Fig. 7. mAP analysis for IITDII and BIPLab showcase the performance of each backbone.

may emerge across different datasets. This analysis will help determine which backbones are better suited for handling various levels of abstraction in sketch-based ear recognition tasks.

In the AMI Backbone Analysis (see Fig. 6(a)), the simplest abstraction for DenseNet121 starts at 26.1% mAP for 16 and increases moderately across the more straightforward abstraction levels (32, 64, 128, 256), reaching 36.8% when considering 256 strokes. However, the combined configurations significantly boost, with DenseNet121 peaking at 40.4% for Config4. EfficientNet models show similar trends, where simple abstraction levels hover around 26.7% to 30.2%, but mAP jumps to approximately 35.3% and 39.2% in combined configurations. InceptionV3 benefits greatly from combined abstractions, with a slight gain in simple levels (reaching 28.3%) but peaking at 31.9% in the final combined configuration. This highlights the strength of combined abstraction levels over individual levels.

In the AWE Backbone Analysis (see Fig. 6(b)), the higher abstraction level, 16, begins with DenseNet121 at 24.8% mAP, gradually increasing through configurations 32, 64, 128, and 256, reaching a modest 27.5% when 256 strokes are considered. However, the combined abstraction levels (Config1 through Config4) exhibit more substantial gains, with DenseNet121 peaking at 33.8% for the most complex configuration, Config4. Similar trends are observed with EfficientNetB0 and EfficientNetB7, which show moderate increases from simpler abstractions (peaking at 28.3% and 27.2%, respectively) but improve significantly in combined configurations, reaching mAP values of around 29.3% and 28.4%. MobileNet also follows this trend, showing minor improvements in simple abstractions but jumping to 32.2% in a complex configuration. Overall, combined abstractions significantly boost mAP compared to individual abstraction levels.

The IITDII Backbone Analysis reveals a similar pattern (see Fig. 7(a)), where simple abstraction levels for DenseNet121 start at 24.8% and incrementally improve across 32, 64, 128, and 256, reaching 27.5%. However, the combined configurations substantially increase mAP, with DenseNet121 reaching 33.8%. EfficientNetB0 and B7 show moderate improvements from 26.7% to 27.2% in simple abstraction levels but achieve higher mAP values in combined configurations, peaking at 29.1% and 29.6%. ResNet152, like the other datasets, sees limited improvements in simple abstraction but jumps to 26.2% in the final combined configuration. The combined configurations consistently outperformed the simple ones.

In the BIPLab Backbone Analysis (see Fig. 7(b)), we see that the simplest abstraction (16) for DenseNet121 starts at 25.5% and gradually improves through 32, 64, 128, and 256, peaking at 35.0%. However, when switching to combined abstractions, the mAP improves significantly, reaching 39.2% at the Config3 complex level. EfficientNetB0 and B7 show consistent trends, with simple abstraction mAP values ranging from 29.7% to 31.9%, but combined configurations yield notable improvements, peaking at 32.6% and 37.4%. InceptionV3 and MobileNet also follow the same pattern, where gains are limited in simple abstractions but increase considerably in combined levels, with MobileNet achieving an impressive 36.6% mAP in the final configuration.

The effect of stroke count on recognition performance is not strictly linear, as increasing the number of strokes does not constantly improve mAP. As the number of strokes grows, the sketch representation gains detail, potentially enhancing recognizability up to a certain point. However, beyond this threshold, additional strokes may introduce extraneous information that aligns differently from the discriminative features most beneficial for effective recognition. This phenomenon can lead to plateaus or even declines in mAP as stroke complexity increases. Furthermore, different backbone architectures, such as DenseNet, ResNet, and MobileNet, respond uniquely to variations in stroke complexity due to their inherent architectural properties and representational capacities. While some architectures are optimized for high-level abstractions, others perform better with more detailed, fine-grained representations, resulting in non-linear mAP trends across varying levels of abstraction.

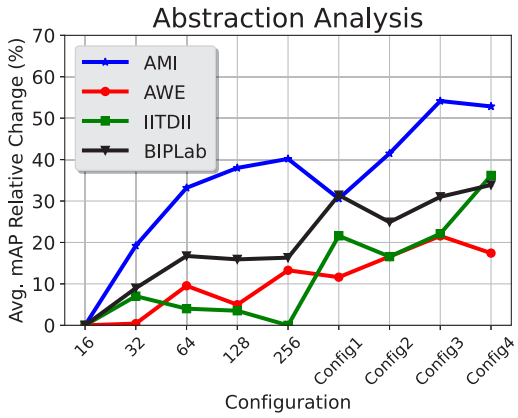
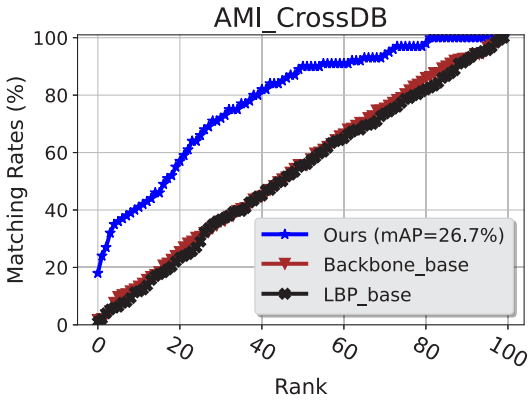


Fig. 8. Relative backbones-average mAP changes across different configurations compared to the first configuration using 16 strokes. Config1, Config2, Config3, and Config4 correspond to 16∪32∪64, 16∪64∪256, 64∪128∪256, and 16∪32∪64∪128∪256, respectively.

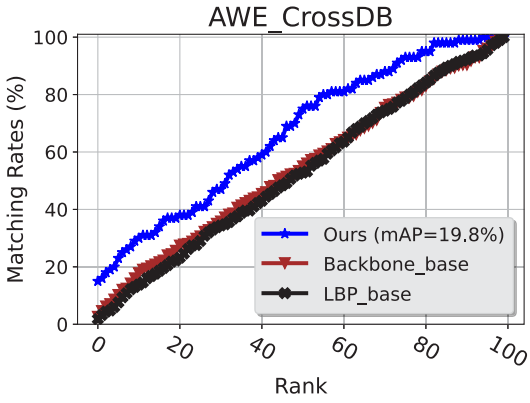
Additionally, with excessive stroke detail, certain backbones may begin to overfit to these finer nuances, potentially reducing their generalization capability. In scenarios where a higher level of abstraction captures the subject's core features effectively. Further, increasing stroke count may inadvertently emphasize irrelevant details, diminishing overall performance. Across all four datasets—AWE, AMI, BIPLab, and IITDII—the pattern is clear: simple abstraction levels (*i.e.*, 16, 32, 64, 128, 256) result in modest mAP improvements for all backbones, with gains generally plateauing around 256. However, performance significantly increases when switching to combined abstraction levels (Config1, Config2, Config3, Config4). DenseNet121 consistently benefits the most, while models like EfficientNetB0, EfficientNetB7, and InceptionV3 also see notable improvements, especially in the combined configurations. This consistent trend highlights the importance of leveraging combined abstraction levels to maximize model performance across all datasets.

6.2. Datasets vs abstraction levels

Fig. 8 represents the percentage change in average mAP for various configurations when analyzing all the different backbones in the context of abstraction levels across the four distinct datasets: AMI, AWE, IITDII, and BIPLab. For the AMI dataset, the increase in mAP is relatively steady, showing a peak performance at the eighth configuration (Config3: 64∪128∪256), which indicates that combining these particular stroke levels yields the best result in terms of mAP improvement relative to the first data point. The AWE dataset shows a more varied trend, with the highest mAP increase observed in the seventh configuration (Config2: 16∪64∪256), suggesting that this combination of strokes is most effective for the AWE dataset. However, in contrast to the overall trend observed across all backbone models, the DenseNet121 emerges as the top performer for the AWE dataset, particularly when considering the eighth configuration (Config3), as demonstrated in Table 2. Interestingly, the IITDII dataset, which focuses on grayscale images of Indian individuals, presents a different pattern, with the most significant mAP increase in the final configuration (Config4: 16∪32∪64∪128∪256), highlighting that a broader range of stroke levels contributes significantly to performance improvement. Lastly, the BIPLab dataset, simulating ear regions during phone calls, demonstrates consistent improvement across configurations, with the highest increase in mAP in the final configuration (Config4). However, similar to the AWE dataset, the top-performing model for the BIPLab dataset deviates from the overall trend observed across all models. Specifically, the optimal performance is achieved when considering the eighth configuration (Config3), as demonstrated in Table 1.



(a) CMC analysis of AMI with baseline methods.



(b) CMC analysis of AWE with baseline methods.

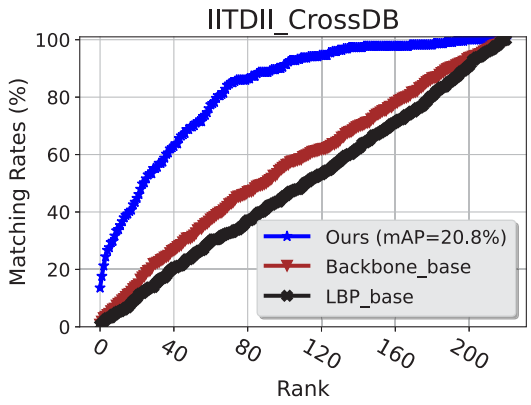
Fig. 9. The CMC curves for AMI and AWE when considering a cross-dataset approach.

These observations underscore the importance of context-specific configuration choices in optimizing mAP improvements. The data reveal that while some datasets benefit from a broad combination of stroke levels, others achieve optimal results with more targeted selections, emphasizing the necessity of tailoring feature abstraction levels to the specific characteristics and challenges of each dataset.

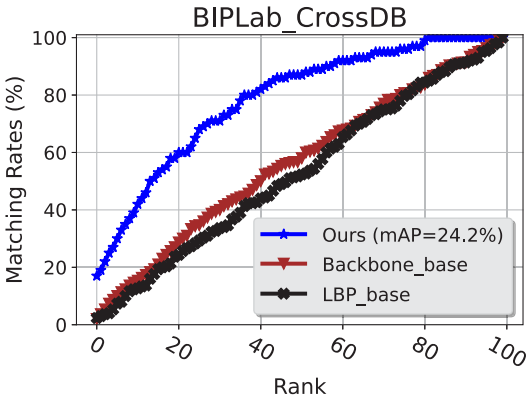
6.3. Cross-dataset performance

The cross-dataset experiment evaluates the generalizability of our approach compared to baseline methods by testing the models on datasets that differ from the training data. This scenario simulates real-world conditions where models encounter variability in subject demographics, image quality, and environmental conditions. Each experiment uses the DenseNet121 backbone, which has been identified as the best-performing backbone and incorporates the optimal abstraction level configuration as reported in Section 5. A single dataset was used exclusively for testing for each experiment, while the remaining datasets were combined and used for training. For instance, if the tested dataset was AMI, then IITDII, AWE, and BIPLab were used for training. The CMC curves presented in Figs. 9(a), 9(b), 10(a), and 10(b) depict the performance of our model, labeled as "Ours", alongside the baseline algorithms $Backbone_{base}$ and LBP_{base} on the AMI, AWE, BIPLab, and IITDII datasets, respectively.

Fig. 9(a) shows the CMC curve for the AMI dataset in the cross-dataset experiment. Our method achieves an mAP of 26.7%, significantly outperforming both $Backbone_{base}$ and LBP_{base} across all ranks. This superior performance indicates the robustness of our approach in adapting to variations in the AMI dataset when trained on different data distributions. In the AWE dataset, as shown in Fig. 9(b), our



(a) CMC analysis of IITDII with baseline methods.



(b) CMC analysis of BIPLab with baseline methods.

Fig. 10. CMC curves for IITDII and BIPLab when considering a cross-dataset approach.

model achieves an mAP of 19.8%. Despite the inherent challenges of the AWE dataset, which includes diverse and in-the-wild ear images, our approach consistently surpasses the baseline models throughout the ranks, demonstrating its capability to generalize well under these unconstrained conditions.

As depicted in Fig. 10(a), our approach reaches an mAP of 20.8% on the IITDII dataset, maintaining a clear lead over the baseline methods. The IITDII dataset, characterized by grayscale images of Indian individuals, presents unique demographic-specific challenges. Finally, the CMC curve for the BIPLab dataset in Fig. 10(b) illustrates that our method attains an mAP of 24.2%, also outperforming the $Backbone_{base}$ and LBP_{base} methods. The BIPLab dataset, which mimics ear images captured during phone calls with realistic conditions such as blurring and limited pose variations, further highlights our model's adaptability and effectiveness in handling real-world scenarios.

When comparing the results of the cross-dataset experiment with the cross-fold validation experiment, it is evident that our model performs better in the cross-fold validation setup, as indicated by higher mAP and rank-1 accuracy scores across all datasets. This discrepancy arises primarily due to the differences in training and testing conditions between the two experiments. In cross-fold validation, the training and testing data come from the same dataset, allowing the model to learn and adapt to the specific characteristics and distributions of that dataset. This setup typically leads to higher performance metrics, as the model encounters less variability and can leverage the consistency within the data folds. However, in the cross-dataset experiment, the model is trained on one dataset but tested on a completely different dataset, which introduces challenges like dataset variability, unseen data distributions, and domain shift. Each dataset has distinct characteristics, such as differences in image quality, resolution, lighting conditions, and subject demographics. Second, the cross-dataset setting exposes the model to data distributions it has not encountered during training. This often includes variations in ear shapes, orientations, and environmental factors that the model has not been specifically trained to handle, resulting in a performance drop. Finally, the inherent domain shift between datasets, such as the difference between controlled environment images in IITDII and the in-the-wild conditions of AWE, further exacerbates the model's ability to maintain high performance, particularly at rank-1, where precise matches are critical.

7. Conclusions

This paper addressed the *sketch-2-image* matching problem in ear data, focusing on the sketch abstraction level. This can be seen as a novelty in biometric recognition, as the previous works in this scope use exclusively the *face* as trait. Hence, ear sketches not only broaden the scope of application for sketch-based biometrics but also tap into a rich vein of biometric data that has remained largely under-exploited. Ear sketches, with their unique contours and features, present a fresh domain for deep learning models to demonstrate their adaptability and effectiveness.

The proposed approach integrates a novel adaptation of triplet loss to handle multiple abstraction levels in ear sketches, represents an advancement in the field. By training our model to recognize and interpret various levels of abstraction — where each level corresponds to a different stroke count — we enable the system to extract and learn from the essential features that define sketches, regardless of their complexity/detail. This methodology enhances the model's robustness and ability to generalize from limited information, a key advantage when dealing with sparse and abstract inputs like sketches.

When comparing to various well-known deep learning architectures, the consistently highest performance of DenseNet121 across all datasets suggests its robustness and adaptability to varying conditions and combining abstraction levels, making it an ideal candidate for sketch-based biometric recognition tasks. Furthermore, the variation in mAP and rank-1 scores across different models and datasets underscores the nuanced nature of sketch-based recognition, where model architecture, dataset characteristics, and abstraction levels play crucial roles in determining performance outcomes.

CRediT authorship contribution statement

David Freire-Obregón: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Joao Neves:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Žiga Emeršič:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization. **Blaž Meden:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology. **Modesto Castrillón-Santana:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Formal analysis, Conceptualization. **Hugo Proença:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partially funded by the Spanish Ministry of Science and Innovation under project PID2021-122402OB-C22 and by the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04. The work due to Hugo Proença was funded by FCT/MEC through national funds and co-funded by FEDER - PT2020 partnership agreement under the projects UIDB/50008/2020, POCI-01-0247-FEDER-033395. This work is also supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP.

Data availability

Data will be made available on request.

References

[1] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 1955–1967, URL https://api.semanticscholar.org/CorpusID:17485723.
[2] H.M. Vazquez, F. Becerra-Riera, A. Morales-González, L. López-Avila, M. Tistarelli, Local deep features for composite face sketch recognition, in: 2019 7th International Workshop on Biometrics and Forensics, IWBF, 2019, pp. 1–6.
[3] CBS News, Police release new 3-d sketch, 911 audio of suspect in lane bryant murders, 2018, https://www.cbsnews.com/chicago/news/tinley-park-lane-bryant-murders-new-sketch/. (Accessed 26 March 2024).
[4] M. Hennessey, Be-lo killer remains unidentified, 2013, https://wcti12.com/archive/be-lo-killer-remains-unidentified. (Accessed 29 March 2024).
[5] Lancashire Police, Images released in search to identify man found dead in rivington, 2024, https://www.lancashiretelegraph.co.uk/news/24180270.images-released-man-found-dead-rivington-last-year/, (Accessed 26 March 2024).
[6] C. Galea, R.A. Farrugia, Forensic face photo-sketch recognition using a deep learning-based architecture, IEEE Signal Process. Lett. 24 (2017) 1586–1590.
[7] S.M. Iranmanesh, A. Dabouei, H. Kazemi, N.M. Nasrabadi, Deep cross polarimetric thermal-to-visible face recognition, in: 2018 International Conference on Biometrics, ICB, 2018, pp. 166–173.
[8] H. Kazemi, S. Soleymani, A. Dabouei, S.M. Iranmanesh, N.M. Nasrabadi, Attribute-centered loss for soft-biometrics guided face sketch-photo recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2018, pp. 612–6128.
[9] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, A. Majumdar, Face sketch matching via coupled deep transform learning, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 5429–5438.
[10] B. Klare, Z. Li, A.K. Jain, Matching forensic sketches to mug shot photos, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 639–646.
[11] L. López-Avila, Y.P. Calaña, Y. Martínez-Díaz, H.M. Vazquez, On the use of pre-trained neural networks for different face recognition tasks, in: Iberoamerican Congress on Pattern Recognition, 2017, URL https://api.semanticscholar.org/CorpusID:6556516.
[12] H. Han, B. Klare, K. Bonnen, A.K. Jain, Matching composite sketches to face photos: A component-based approach, IEEE Trans. Inf. Forensics Secur. 8 (2013) 191–204.
[13] D. Liu, J. Li, N. Wang, C. Peng, X. Gao, Composite components-based face sketch recognition, Neurocomputing 302 (2018) 46–54.
[14] Y. Song, Z. Zhang, H. Qi, r-btn: Cross-domain face composite and synthesis from limited facial patches, in: AAAI Conference on Artificial Intelligence, 2017.
[15] L. Meijerman, A. Thean, G. Maat, Earprints in forensic investigations, Forensic Sci. Med. Pathol. 1 (4) (2005) 247–256.
[16] Y. Vinker, E. Pajouheshgar, J.Y. Bo, R.C. Bachmann, A.H. Bermano, D. Cohen-Or, A. Zamir, A. Shamir, CLIPasso: Semantically-aware object sketching, ACM Trans. Graph. 41 (4) (2022).
[17] A. Das, Y. Yang, T.M. Hospedales, T. Xiang, Y.-Z. Song, Bézier sketch: A generative model for scalable vector sketches, in: European Conference on Computer Vision, 2020.
[18] S. Ge, V. Goswami, L. Zitnick, D. Parikh, Creative sketch generation, in: International Conference on Learning Representations, 2021.
[19] B. Klare, A.K. Jain, Sketch-to-photo matching: a feature-based approach, in: Defense Commercial Sensing, 2010.
[20] H.S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, Memetically optimized mcwld for matching sketches with digital face images, IEEE Trans. Inf. Forensics Secur. 7 (2012) 1522–1535.
[21] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 513–520.
[22] B. Klare, A.K. Jain, Heterogeneous face recognition using kernel prototype similarities, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1410–1422.
[23] P. Mittal, M. Vatsa, R. Singh, Composite sketch recognition via deep network - a transfer learning approach, in: 2015 International Conference on Biometrics, ICB, 2015, pp. 251–256.
[24] K. Chang, K. Bowyer, S. Sarkar, B. Victor, Comparison and combination of ear and face images in appearance-based biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1160–1165.
[25] S. Barra, M. De Marsico, M. Nappi, D. Riccio, Unconstrained ear processing: What is possible and what must be done, in: Signal and Image Processing for Biometrics, 2014, pp. 129–190.
[26] Ž. Emeršič, V. Štruc, P. Peer, Ear recognition: More than a survey, Neurocomputing 255 (2017) 26–39.
[27] M. Burge, W. Burger, Ear biometrics, biometrics: Personal identification in networked society, 1999, pp. 273–285.
[28] D. Hurley, M. Nixon, J. Carter, Automatic ear recognition by force field transformations, in: IEE Colloquium on Visual Biometrics (Ref.No. 2000/018), 2000, pp. 7/1–7/5.
[29] J. Bustard, M. Nixon, Toward unconstrained ear recognition from two-dimensional images, IEEE Trans. Syst. Man Cybern. Syst. Humans 40 (2010) 486–494.
[30] A. Kumar, T. Chan, Robust ear identification using sparse representation of local texture descriptors, Pattern Recognit. 46 (2013) 73–85.
[31] H. Alshazly, C. Linse, E. Barth, T. Martinetz, Deep convolutional neural networks for unconstrained ear recognition, IEEE Access 8 (2020) 170295–170310.
[32] D. Freire-Obregón, M.D. Marsico, P. Barra, J. Lorenzo-Navarro, M. Castrillón-Santana, Zero-shot Ear Cross-dataset Transfer for Person Recognition on Mobile Devices, Pattern Recognit. Lett. 166 (2023) 143–150, http://dx.doi.org/10.1016/j.patrec.2023.01.012.
[33] A. Abate, M. Nappi, S. Ricciardi, I-Am: Implicitly authenticate Me—Person authentication on mobile devices through ear shape and arm gesture, IEEE Trans. Syst. Man Cybern. Syst. 49 (2019) 469–481.
[34] S. Koley, A.K. Bhunia, A. Sain, P.N. Chowdhury, T. Xiang, Y.-Z. Song, How to handle sketch-abstraction in sketch-based image retrieval? in: 2024 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2024.
[35] K. Mukherjee, H. Huey, X. Lu, Y. Vinker, R. Aguina-Kang, A. Shamir, J.E. Fan, Seva: Leveraging sketches to evaluate alignment between human and machine visual abstraction, 2023, arXiv abs/2312.03035.
[36] A. Sain, A.K. Bhunia, Y. Yang, T. Xiang, Y.-Z. Song, Stylemeup: Towards style-agnostic sketch-based image retrieval, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 8500–8509.
[37] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T.M. Hospedales, C.C. Loy, Sketch me that shoe, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 799–807.
[38] W. Wan, Y. Gao, H.J. Lee, Transfer deep feature learning for face sketch recognition, Neural Comput. Appl. 31 (12) (2019) 9175–9184.
[39] Y. Zhao, W. Wang, H. Zhang, B. Hu, Learning homogeneous and heterogeneous co-occurrences for unsupervised cross-modal retrieval, in: 2021 IEEE International Conference on Multimedia and Expo, ICME, 2021, pp. 1–6.
[40] Y. Qi, Y.-Z. Song, H. Zhang, J. Liu, Sketch-based image retrieval via siamese convolutional neural network, in: 2016 IEEE International Conference on Image Processing, ICIP, 2016, pp. 2460–2464.
[41] F. Wang, L. Kang, Y. Li, Sketch-based 3d shape retrieval using convolutional neural networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1875–1883.
[42] E. González-Sánchez, Biometría de la oreja (Ph.D. thesis), Universidad de Las Palmas de Gran Canaria, 2008.
[43] A. Kumar, C. Wu, Automated human identification using ear imaging, Pattern Recognit. 45 (2012) 956–968.
[44] A. Buslaev, V. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2020).
[45] H. Alshazly, C. Linse, E. Barth, S.A. Idris, T. Martinetz, Towards explainable ear recognition systems using deep residual networks, IEEE Access 9 (2021) 122254–122273.
[46] S. El-Naggar, T. Bourlai, Exploring deep learning ear recognition in thermal images, IEEE Trans. Biom. Behav. Identity Sci. 5 (1) (2023) 64–75.
[47] Y. Khaldi, A. Benzaoui, A. Ouahabi, S. Jacques, A. Taleb-Ahmed, Ear recognition based on deep unsupervised active learning, IEEE Sens. J. 21 (18) (2021) 20704–20713.
[48] Ž. Emeršič, A. Kumar, B. Harish, W. Gutfeter, A. Pacut, E. Hansley, M. Pamplona Segundo, S. Sarkar, H. Park, V. Štruc, Štruc the unconstrained ear recognition challenge 2019, in: 2019 International Conference on Biometrics, ICB, 2019, pp. 1–5.
[49] Ž. Emeršić, T. Ohki, M. Akasaka, T. Arakawa, S. Maeda, M. Okano, Y. Sato, A. George, S. Marcel, I. Ganapathi, et al., The unconstrained ear recognition challenge 2023: Maximizing performance and minimizing bias, in: 2023 IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2023, pp. 1–10.