Pattern Recognition 179 (2026) 113711

Contents lists available at ScienceDirect

Pattern Recognition

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

Bones to identity: Generative contrastive fusion for cross-modality medical person identification from skeletal data☆

Chaoqun Niu [a,b], Dongdong Chen [c], Jizhe Zhou [a,b], Jian Wang [a,b], Quan-Hui Liu [a,b], Caiyang Yu [a,b], Yuan Li [d], Wei Ju [a,b], Jiancheng Lv [a,b,*]

[a] College of Computer Science, Sichuan University, Chengdu, 610065, Sichuan, China
[b] Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, 610065, Sichuan, China
[c] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK
[d] West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, 610065, Sichuan, China

ARTICLE INFO

Dataset link: Neural-Boneprint

Keywords:
Medical person identification
Cross modal
Skeletal data
Generative contrastive learning

ABSTRACT

Forensic person identification is critical in accidents and criminal investigations. Existing methods based on soft tissue or DNA can be unavailable if the body is skeletonized or charred. Fortunately, bones persist for a long time, raising a natural question: ***can we identify a person using bone data?*** To address this, we consider two fundamental questions: First, which skeletal data should be utilized? Second, what features should be extracted for identification, and how? We propose ***Neural Boneprint*** (NBP), a novel biometric-like identifier fused from cross-modality medical skeletal data, for personal identification. Specifically, we exploit the thoracic skeletal data, including chest radiographs (CXRs) and volume-rendered CT (VRT), as an example to explore the viability of the NBP. We present two complementary fusion paradigms grounded in generative contrastive learning: (i) an explicit fusion approach that translates between modalities and fuses real/synthetic pairs via a dual-reconstruction contrastive network; and (ii) an implicit fusion approach that integrates CXR and VRT directly within a diffusion-based generative process, where cross-modal contrastive constraints enforce identity consistency without explicit translation or feature concatenation. Both yield a modality-invariant latent embedding, the NBP, which captures person-specific skeletal information while suppressing modality-specific artifacts. Evaluated on real clinical data, our framework achieves 90.49% Rank-50 accuracy, substantially outperforming prior art. Further experiments with up to 10,000 distractors demonstrate robustness and scalability. By unifying explicit and implicit fusion under a common generative contrastive framework, this work advances the theory and practice of cross-modality information fusion in medical identity verification. The code is available at *Neural-Boneprint*.

1. Introduction

Person identification is critical in forensic investigations involving accidents, disasters, or criminal cases. Historically, scholars have delved into examining diverse forms of biological evidence to facilitate personal identification and authentication. This exploration encompasses a range of biometric indicators, such as deoxyribonucleic acid (DNA) [1], faces [2], fingerprints [3], etc.

Nevertheless, in forensics, the identification of a body or corpse that has undergone extensive decomposition, intentional mutilation, or incineration presents significant challenges. Soft tissue markers, such as facial features and fingerprints, frequently become non-viable or unobtainable for identification under such conditions. Moreover, the extraction of DNA from these compromised remains is markedly difficult due to the progressive degradation of genetic material over time [4]. Beyond the inherent technical complexities, the identification process is further constrained by financial considerations, temporal requirements, and the comprehensiveness of existing DNA databases. The efficacy of DNA analysis is contingent upon the presence of the individual's DNA sequence, or that of close relatives, within the relevant databases. Absent prior sequencing and archiving, the identification process encounters substantial impediments, illustrating the limitations of current forensic methodologies in certain scenarios.

Forensic studies have demonstrated the feasibility of manually identifying individuals by comparing ante- and post-mortem skeletal imaging materials [5,6]. However, these analyses typically rely on the

☆ This article is part of a Special issue entitled: 'PR_Evolving Multi-View Learning' published in Pattern Recognition.

* Corresponding author.
E-mail address: lvjiancheng@scu.edu.cn (J. Lv).

https://doi.org/10.1016/j.patcog.2026.113711
Received 22 December 2025; Received in revised form 18 March 2026; Accepted 10 April 2026
Available online 20 April 2026

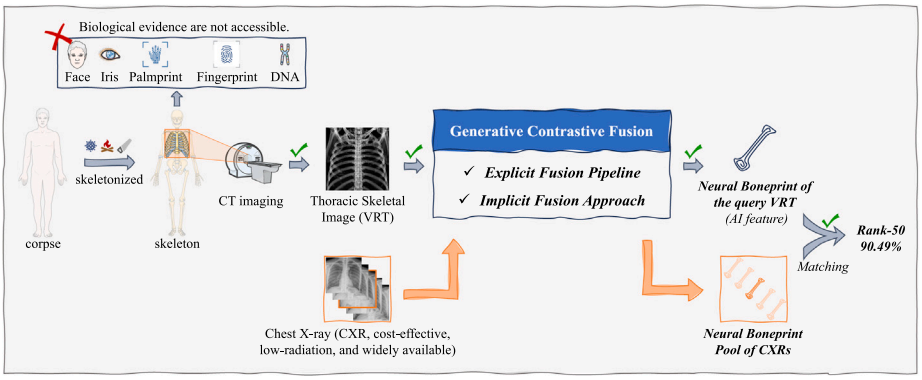

Fig. 1. When a corpse is deeply skeletonized (highly decomposed, burned, or deliberately destroyed), biological evidence involving soft tissue or DNA is often unavailable. We present ***Neural Boneprints*** for person identification by taking the thoracic skeleton as an example.

expertise of forensic professionals. This naturally raises the question: ***Is it possible to develop methodologies for direct identification of individuals based solely on skeletal data?*** To achieve this, two fundamental questions must be addressed. First, which skeletal data should be utilized? Second, what features should be extracted for identification, and how?

Modern hospitals routinely archive diverse skeletal imaging data, such as low-cost, widely available chest X-rays (CXRs) and high-fidelity volume-rendered CT (VRT) scans, creating a rich but heterogeneous repository. The core challenge lies not merely in extracting features, but in fusing information across these disparate modalities to form a coherent identity representation. Unlike standard multimodal tasks where sensors capture aligned views [7], CXR and VRT differ drastically in anatomy visibility (soft tissue overlap vs. pure bone), acquisition posture (standing vs. supine), and imaging physics—rendering pixel-level or early-feature fusion ineffective.

To this end, we advocate a fusion-centric perspective on medical person identification and formulate this task as a latent-level fusion problem via generative contrastive learning, where heterogeneous medical observations are projected into a unified identity-consistent representation space. We introduce the Neural Boneprint (NBP), a learned identity embedding derived through cross-modality information fusion, and present two novel instantiations within a unified Generative Contrastive Fusion (GCF) framework: (i) an explicit fusion pipeline combining modality translation with latent alignment, and (ii) an implicit fusion approach achieved through joint generative contrastive modeling, as illustrated in Fig. 1. Experimental results on clinical data demonstrate the effectiveness of NBP for forensic person identification. In summary, this work makes the following contributions:

(i) We propose a novel perspective on person identification with medical image data by introducing NBP, a learned identifier derived from skeletal images. NBP serves as a biometric-like representation that enables effective cross-modality person identification, constituting the first solution to medical person re-identification across imaging modalities, to our knowledge.
(ii) A novel generative contrastive deep learning framework with two fusion-centric implementation approaches (explicit and implicit fusion) is presented to extract NBPs from CXR and VRT images of the thoracic skeleton. NBP is architecture agnostic, allowing its extraction using different networks.
(iii) Experimental results on real clinical data demonstrate the effectiveness of NBP in identifying people, achieving a Rank-50 identification accuracy of 90.49%, significantly surpassing existing methods.

Extension Statement. A preliminary version of this work was published in ACM MM [8]. The journal extension substantially expands upon the original conference paper in several key aspects.

(i) **New Method.** We propose a novel diffusion-based generative contrastive learning approach that performs implicit fusion within a unified generative process, eliminating the need for explicit modality translation. This one-stage method is conceptually simpler and still highly effective, improving Rank-50 accuracy from 84.79% (conference version) to 86.69%.
(ii) **Theoretical Consistency and Framework Generalization.** The journal version provides a more complete formulation of our proposed GCF paradigm, demonstrating how both the explicit and implicit approaches serve as theoretically unified instantiations of the same fusion principle for solving this interesting but challenging task.
(iii) **Enhanced Explicit Fusion Performance.** The explicit fusion pipeline is significantly improved through refined loss balancing, architectural tuning, and extensive experiments. The updated version achieves a Rank-50 accuracy of 90.49%, far surpassing the conference results.
(iv) **Large-Scale Robustness Evaluation.** We introduce a new robustness benchmark using up to 10,000 distractor CXRs from the ChestX-ray8 [9], representing a substantially more realistic and scalable evaluation setting than the small-scale distractor setup in the conference paper. Our method demonstrates strong scalability under this challenging evaluation.
(v) **Additional Analyses, Visualizations, and Baselines.** We provide deeper discussions of human expert criteria in forensic skeletal comparison, introduce several new baselines, and supply extensive qualitative visualizations, confirming alignment with forensic expert practices and providing deeper interpretability.

2. Related work

Clinical imaging is frequently utilized to retrieve target categories or verify patient identity [10,11] for robust medical data archiving. Additionally, comparing antemortem and postmortem imaging is widely applied in the identification of human remains [12].

Traditionally, forensic identification relies heavily on the manual comparison of antemortem and postmortem records. This includes analyzing dental radiographs [13], utilizing head and neck CT/MRI scans [6], and visually assessing full CT images [12], though the requisite number of concordant traits lacks consensus. Other manual approaches examine specific skeletal structures like clavicles and vertebrae [14] or fuse postmortem CTs with antemortem chest radiographs (CXRs) [15]. Ultimately, these manual methods are labor-intensive, require extensive domain expertise, and are challenging to deploy on a large scale. Recent automatic methods attempt to address these limitations but remain inadequate. Early cross-modal matching between antemortem CXRs and postmortem CTs using hand-crafted features [16] or Discrete Fourier Transform [17] yields unsatisfactory performance and relies on extremely small datasets.

Concurrently, CXR re-identification has gained traction for patient archival purposes [10,11,18]. While identifying individuals solely through CXRs appears straightforward, existing methods rely heavily on the presence of soft tissues and internal organs, rendering them unsuitable for skeletal matching. In forensic scenarios, postmortem organ decomposition severely disrupts visual comparisons, and for fully skeletonized remains, organ-based features become entirely invalid and misleading.

3. Motivation

In this paper, we aim to explore a seldom-exploited yet fundamental observation: *bones typically endure for an extended period, either within corpses or skeletons.* We assume that there is an implicit feature *boneprint* in bones, similar to palmprints and fingerprints, which encodes identity information and is widely present in skeletons and skeletal data.

3.1. Which bones? Thoracic skeleton

It is worth noting that not all bones on the skeleton can be a suitable candidate for large-scale person identification. Those available bones, such as vertebrae and skulls [14,19], contain identifiable boundaries and rich morphology that varies from person to person. In brief, efficient skeleton-based person identification requires the skeletal data (i) to contain identifiable boundaries and distinct morphology, and (ii) to be easily collected and organized in a large-scale matching pool.

While it is commonly acknowledged that human faces and skulls contain identifiable features [19], we seek to investigate overlooked aspects of the skeletal structure that may also hold valuable identity information, and the thoracic skeleton serves as an apt example for this purpose. Thoracic skeletons, comprising ribs, vertebrae, and sternums [20,21], have been employed as manual comparison materials to estimate sex and age due to complex morphology and distinct visual individual differences. Consequently, this paper will explore the thoracic skeleton data to learn thoracic boneprints, thereby expanding the toolkit available for forensic identification.

3.2. Which thoracic skeletal data? VRT & CXR

Forensic pathologists are often required to preprocess computed tomography (CT) images by volume rendering technology (VRT) [22] to obtain VRT images for analytical studies. For the corpse, chest VRT enables the generation of a clear thoracic skeleton, thus avoiding the potential confounding effects of soft tissue and organ decomposition on skeletal observation. However, it is impractical to construct an adequate pool from VRT images. That is, not everyone has had a CT scan because the use of CT increases the risk of radiation-induced cancer [23], resulting in a paucity of VRT images. As a result, it is unlikely that there is a pre-stored CT image to compare with a query VRT image of an unnamed corpse.

Fortunately, the chest X-ray (CXR) is a routine and cost-effective component of the physical examination. For example, in industrialized countries, about 238 CXRs are acquired annually per 1000 people [24,25]; in the U.S. alone, 129 million CXRs were taken in 2006 [26]. The CXR is often the initial imaging study acquired and remains pivotal to the screening, diagnosis, and management of a broad range of conditions, with numbers likely higher post-COVID-19.

More importantly, CXR images typically contain all the skeletal elements [27] of the chest and their identities. This illustrates that CXRs can constitute a substantial repository for the identification of thoracic skeletons. Simply using CXRs or synthetic CXRs of corpses may be proposed. However, for corpses that are not fully decomposed, the decomposition and expansion of organs and soft tissues can significantly impact observation and comparison. Furthermore, for skeletonized corpses, the simulated soft tissues and organs in synthetic CXRs can be misleading and thus unsuitable for analysis. The question then becomes: How to exploit the vast amount of CXR data and introduce its rich identification information into the small VRT data? Can we learn boneprint from CXR and VRT images, and how?

3.3. Where do forensic experts focus on the CXR and VRT images? Morphology & boundary

CXRs contain more anatomical structure overlaps, while VRT images only contain skeletons. Manual methods comparing these imaging materials from antemortem and postmortem [5,12,14] are usually focused on the overall skeletal morphology and boundaries, as labeled in Fig. 2(a).

For the same person, some skeletal details are strikingly similar: the boundaries and curvatures of each rib in the VRT image are almost identical to those in the CXR. However, the whole structure is slightly different. This is because the VRT image and the CXR not only belong to different domains but are also taken in different postures and at different times.

When patients take the CXR, they stand up and hold their arms flat with normal breathing. During the VRT image, they lie flat on the plate and raise their arms upward with deep breathing. The different states lead to the deformation of the whole structure of the thoracic skeletons, as shown in Fig. 2(b), which makes the thoracic skeletons in CXRs and VRT images not strictly pixel-level mapping but potentially non-linear stochastic mapping.

Furthermore, VRT images and CXRs from two individuals with similar overall thoracic skeletal morphology are presented in Fig. 3. Although this resemblance may initially obscure contrasts, a closer examination reveals subtle variations in chest contours and rib boundary details. Notably, even within skeletons exhibiting overall similarity, these minute distinctions are present. The disparities can be expected to be more prominent in the thoracic skeletons displaying significant morphological differences.

3.4. Challenges

To answer these questions, several challenges related to skeletal data (VRT, CXR) need to be addressed. The main challenges can be categorized as follows:

(i) **Single Shot.** There is a notable imbalance in the availability of image types, with VRT images being relatively scarce and sparsely distributed compared to the abundant and densely available CXRs. This disparity poses significant challenges for data handling and analysis [28]. Furthermore, our dataset contains only 2 images per category (a single CXR-VRT image pair per individual, single shot), which limits the extraction of reliable boneprint features.

(ii) **Large Intra-Class Variation.** The considerable modality gap between VRT images and CXRs results in significant intra-class variation. CXRs, in particular, exhibit complex overlaps of various anatomical structures, unlike VRT images that predominantly display skeletal features. Additionally, the varied postures during imaging lead to bone deformation, further complicating the analysis.

(iii) **Small Inter-Class Difference.** The subtlety of thoracic skeletal differences between individuals poses a formidable challenge. Unlike other identification tasks where inter-class features are more distinct, the skeletal features in our case are less conspicuous, sometimes barely discernible to the human eye. Compounding by the physiological similarities in rib numbers and orientations shared among humans makes differentiation based on these features challenging.

Overcoming the aforementioned challenges requires sophisticated approaches to image processing, data analysis, and feature extraction. The next section will introduce novel generative contrastive deep-learning approaches to solve these challenges and learn NBPs for identification.

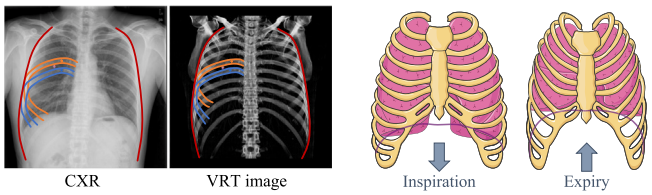

Fig. 2. (a) When comparing the CXR with the VRT image from the same individual, we focus on the overall skeletal morphology (red), the boundaries (orange and blue), the width of ribs (orange and blue bidirectional arrows), and the inter-rib space (orange to blue gradient bidirectional arrows). (b) The different respiratory states lead to the deformation of the thoracic skeletons.

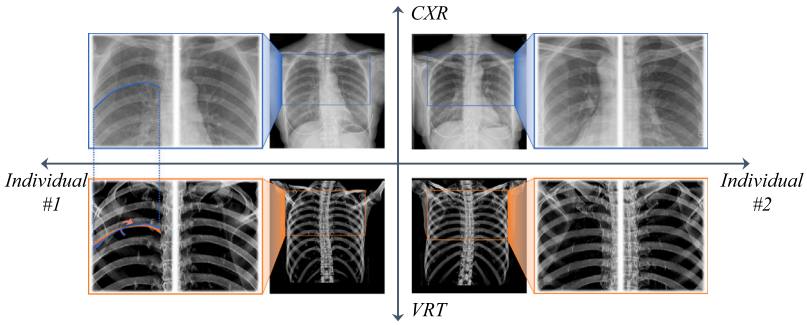

Fig. 3. VRT images and CXRs from two individuals with similar thoracic skeletal morphology. In Individual #1, the fourth left rib (blue solid line) aligns with its VRT counterpart (orange solid line) through rotation alone, consistent with a rigid transformation between CXR and VRT. In contrast, the thoracic skeletons as a whole do not achieve alignment under rigid transformations, exhibiting clear non-rigid discrepancies between the two imaging modalities.

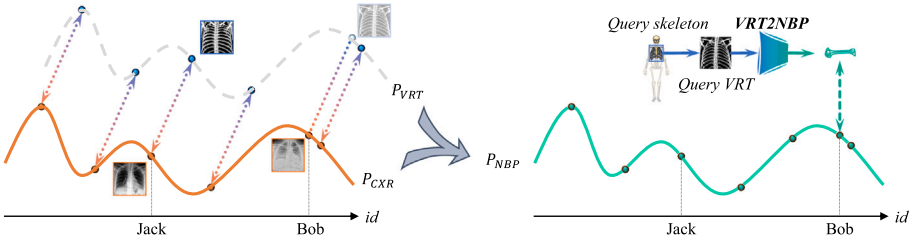

Fig. 4. The VRT images (blue dot) are sparsely distributed (P_{VRT}, gray dashed line) while the CXRs (orange dot) are densely available (P_{CXR}, orange solid line). An explicit generative contrastive approach learns identity-based cross-modal mappings, enabling data complement [29] to obtain the modality-invariant NBP. The VRT2NBP module extracts NBPs from VRT images.
Source: Adapted from [8].

4. NBP learning via generative contrastive fusion

We formulate person identification from skeletal images as a cross-modality information fusion problem, where the goal is to integrate complementary identity cues from CXR and VRT into a single, discriminative representation, the NBP. Recognizing that fusion can occur at different levels of abstraction, we propose two strategies:

(i) **Explicit Fusion**: A two-stage pipeline that bridges the modality gap via bidirectional translation, and fuses real and synthetic views in a shared latent space.
(ii) **Implicit Fusion**: A one-stage end-to-end approach that performs fusion not through architectural modules, but through a unified generative-contrastive objective, where identity consistency across modalities is enforced during diffusion-based synthesis.

Both approaches instantiate the broader principle of GCF: using generation to model modality variation and contrastive learning to distill invariant identity signals. In this paper, let x and y be the VRT image and CXR image.

4.1. Explicit fusion pipeline (two stage method)

In medical examinations and data collection, the number of CXRs is more than that of VRT images due to less radiation and a lower price, which makes a dense CXR distribution and the VRT images being sparsely distributed. Hence, in our two-stage method, we treat those individuals with only one modality as incomplete data [30] and aim to achieve cross-modality identifiable completion. Then, obtain a fused unique feature which unifies identity consistency of both modalities, similar to biometrics but data-driven, named NBP, as illustrated in Fig. 4.

Therefore, in the following sections, we will provide a comprehensive description of our explicit fusion NBP algorithm, as shown in Fig. 5.

It contains three modules: (1) Cross-modality Translation. Images are translated into each other's modality to bridge the modality gap and enhance data completeness. (2) Cross-modality Fusion. A dual reconstruction network with contrastive learning fuses fine-grained representations and optimizes inter- and intra-class distances to extract NBPs. (3) NBP-Bank construction. An NBP Bank is constructed from

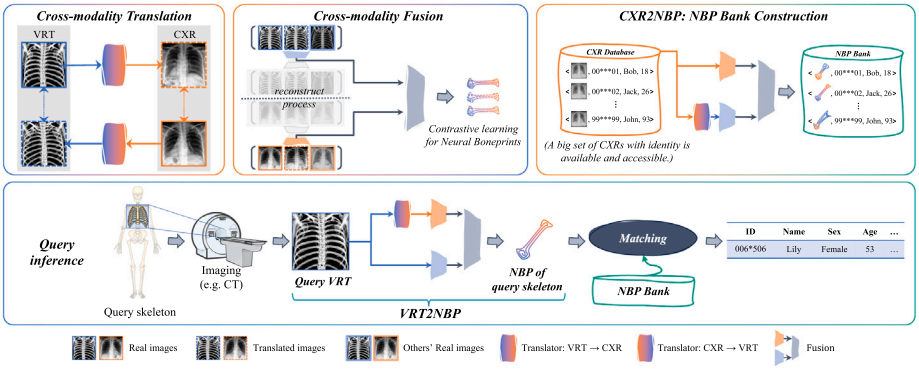

Fig. 5. Explicit fusion pipeline. Top: Query phase. For a query skeletal corpse, we obtain the query NBP from the query skeletal data (VRT image) in the VRT2NBP process. Then match the query NBP with the nearest one in the NBP bank. Bottom: Training phase of the cross-modality translation and cross-modality fusion modules, and construction of the NBP bank via the CXR2NBP process. Translated CXR/VRT denotes synthetic images generated from VRT/CXR, respectively. *Source:* Adapted from the conference version [8].

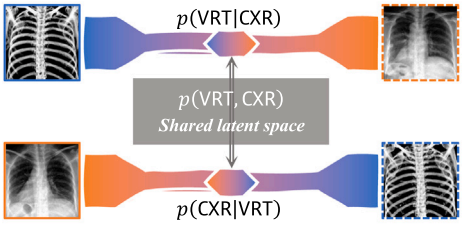

Fig. 6. Model the processes of $p(\mathrm{CXR}|\mathrm{VRT})$ and $p(\mathrm{VRT}|\mathrm{CXR})$ to build the shared latent space.

CXR data. Query VRT images are matched against NBPs in the bank for identification. Two domain translators are denoted as $\mathsf{T}_{xy} : x \to y$ and $\mathsf{T}_{yx} : y \to x$. The algorithm ultimately outputs a NBP through a fusion network F, represented as $\mathsf{F} = \mathrm{NBP}(x, y)$.

4.1.1. Cross-modality translation

The primary obstacle in converting between VRT images and CXRs is the notable differences in anatomical structures and the distortion of the thoracic skeleton. This creates a significant gap between the two modalities that cannot be effectively bridged by straightforward unidirectional translation methods [31].

From a probabilistic view, unidirectional methods focus solely on $p(\mathrm{CXR}|\mathrm{VRT})$ or $p(\mathrm{VRT}|\mathrm{CXR})$, preventing exploration of the intricate semantic relationships between VRT images and CXRs. To address this, we propose a bidirectional process, simultaneously learning from VRT to CXRs and vice versa. By deliberately aligning their intermediate latent variables in terms of distribution, we establish a preliminary joint latent space between VRT and CXRs, denoted as $p(\mathrm{VRT}, \mathrm{CXR})$, as shown in Fig. 6.

Specifically, we employ two transformation networks to model the processes of $p(\mathrm{CXR}|\mathrm{VRT})$ and $p(\mathrm{VRT}|\mathrm{CXR})$ separately. Furthermore, an extra ℓ_2 loss penalty is adopted to maintain fine-grained personally identifiable information. The real VRT image is paired by identity with the CXR image at the semantic level, so the translated VRT image should be identical to the original VRT image due to the individual identity. In this manner, translated images can be ensured to hold the same identity information as the original. It can also be viewed as sampling in the vicinity of the true image in manifold space.

4.1.2. Cross-modality fusion

Following the translation, we obtain an image pair (real, translated) for each individual. Given the design of preserving identity, the translated distribution is already close enough to the real in the identity manifold, allowing us to treat them as equivalent.

We employ a dual-input reconstruction network based on contrastive learning to fuse distinguishable skeletal representations to the latent embedding layer F. It maps each VRT-CXR image pair to a joint latent embedding in a manifold space where identity information is the primary constraint [2]. It contains a VRT encoder–decoder module, a CXR encoder–decoder module, and a latent fusion module. Each encoder–decoder module reconstructs real or translated images to learn latent fine-grained skeletal representations, which are then fused into NBPs. We compute the mean squared error (MSE) between reconstructed and original images in the pixel space to refine better learning of fine-grained identical latent features. We apply contrastive loss [32,33] on the fused embeddings of real, translated, and the other's real image pairs to minimize intra-class distance while maximizing inter-class distance. During training, the encoder–decoder and latent fusion modules are jointly optimized. In the application phase, we retain only encoders and the latent fusion module with weights frozen, discarding the decoders.

4.1.3. NBP-Bank construction

The main idea is that CXRs are widely available and identity-rich, while VRT images are common for skeletal remains. This process aims to develop an identification function $\mathrm{ID}(x)$ for efficient identification using CXR databases. Given a CXR database, we process the following 3 steps:

(i) *Modality Unification:* All CXRs $\{y_i\}$ are first translated into VRT images $\{\mathsf{T}_{yx}(y_i)\}$ using a translator T_{yx}.

(ii) *Joint Embedding Generation:* CXR-VRT image pairs $\{(\mathsf{T}_{yx}(y), y)\}$ are mapped via F to joint latent embeddings (NBPs) that capture cross-modal relationships, incorporating identity information.

(iii) NBP-*Bank Creation:* The NBP-Bank, denoted as $\mathcal{Z}$, is constructed as a searchable table with embeddings as keys and identities as values:

$$\mathcal{Z} \triangleq \{(\mathrm{Key} = \mathsf{F}(\mathsf{T}_{yx}(y_i), y_i), \mathrm{Value} = \mathrm{ID}(y_i))\}. \tag{1}$$

4.2. Implicit fusion approach (one-stage method)

Unlike explicit fusion with translation modules, our one-stage method implicitly integrates modalities via diffusion-based generation and contrastive learning. It trains a diffusion model conditioned on identity to produce modality-consistent samples, while a contrastive loss aligns embeddings of the same person across CXR and VRT in the latent space. Fusion emerges from joint optimization, no explicit translation or concatenation, learning a unified identity manifold that preserves skeletal signatures despite modality-specific distortions.

We thus interpret this approach as implicit cross-modality fusion, where information integration is achieved through shared generative

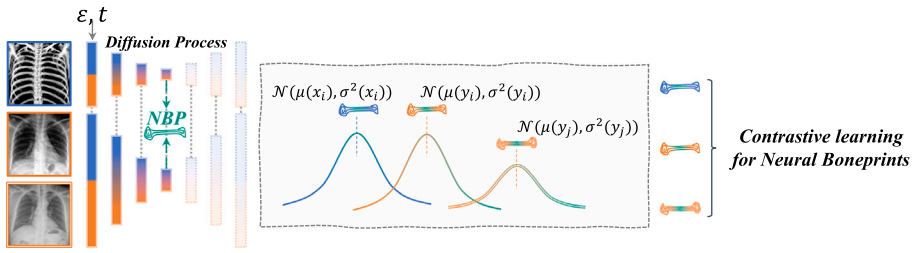

Fig. 7. Leveraging the consistency of skeletal identity information, we model all images of an individual within a single modality, which can be potentially influenced by time t and noise ϵ, as samples from a Gaussian distribution centered at the observed image. Contrastive learning is then applied to optimize feature distances, encouraging maximal similarity among intra-individual samples and minimal similarity across inter-individual ones.

dynamics and contrastive alignment, rather than handcrafted fusion layers.

Ideally, each individual could possess multiple original scans across both modalities. The skeletal biometric identity information, i.e., NBP, involved in these images enables them to follow an individual-specific distribution p. Specifically, while an individual may undergo numerous CXRs, the skeletal identity features remain invariant across these images. If pathological conditions are absent, these images will generally maintain high consistency, with thoracic skeletons exhibiting only minor deformations constrained within the developmental homeostasis and biomechanical constraints.

Therefore, for those individuals who have only one image, if we let y_i represent the collected single CXR of the ith individual, their multiple-shot CXR distribution can be denoted as:

$$p(\hat{y}_i|y_i) = \mathcal{N}(\boldsymbol{\mu}(y_i), \boldsymbol{\sigma}^2(y_i)), \tag{2}$$

where $\hat{y}_i$ represents the other CXRs that have been taken or are potentially not taken. Similarly, for the VRT modality, the distribution is $p(\hat{x}_i|x_i) = \mathcal{N}(\boldsymbol{\mu}(x_i), \boldsymbol{\sigma}^2(x_i))$. Due to the consistency of identity information contained in the images, the image distributions of the same individual across different modalities should be more similar. In contrast, image distributions of different individuals should show greater disparity, even within the same modality. Consequently, the optimization objective should be as follows:

$$\min \sum_{i\neq j}^{m\in\mathcal{M}} [KL(p(\hat{y}_i|y_i)\|p(\hat{x}_i|x_i)) - KL(p(\hat{m}_i|m_i)\|p(\hat{m}_j|m_j))], \tag{3}$$

where $\mathcal{M} = \{x, y\}$, i, j for different individuals. However, the distributions $p(\hat{y}_i|y_i)$ and $p(\hat{x}_i|x_i)$ with their variances are intractable, although we can simply regard y_i, x_i as their means, respectively.

Hence, as illustrated in Fig. 7, we design a one-stage method inspired by the diffusion models [34]. In the diffusion process, each noisy CXR image of step t with added noise ϵ can be regarded as a sampling from the distribution p derived from clean images, which can be formulated as follows:

$$\hat{y}_i^t = \boldsymbol{\mu}(y_i) + \epsilon(t)\boldsymbol{\sigma}^2(y_i), \epsilon \sim \mathcal{N}(0, I). \tag{4}$$

Similarly, for noisy VRT images, it can be formulated as $\hat{x}_i^t = \boldsymbol{\mu}(x_i) + \epsilon(t)\boldsymbol{\sigma}^2(x_i), \epsilon \sim \mathcal{N}(0, I)$. Since both clean and sampled images contain unique identity information, when the encoder–decoder in the diffusion model predicts the noise added at step t, the features output by the encoder E should also encapsulate this unique identity information to form the proposed NBP, i.e., $\mathsf{NBP} = \mathsf{E}(\hat{m}_i^t, t), m \in \mathcal{M}$. In this way, besides the original MSE loss of the noise in the diffusion model, we can adopt the contrastive loss of NBPs from the VRT image, CXR, and others' CXRs to optimize the intra- and inter-class distances as a surrogate objective for maximizing distributional similarity within individuals and disparity across different individuals.

For the construction of NBP Bank $\mathcal{Z}$, we employ only the trained encoder E to obtain NBPs from clean CXR images:

$$\mathcal{Z} \triangleq \{(\mathsf{Key} = \mathsf{E}(y_i, t), \mathsf{Value} = \mathsf{ID}(y_i))\}, \tag{5}$$

where t is discretionary, and the experiments show that t is not significantly correlated with performance.

4.3. Query inference: Identification of persons with NBP

For both fusion approaches, the NBP can be derived from either a CXR or a VRT image, enabling bidirectional retrieval. In the explicit method, we use the appropriate translator to obtain a paired representation; in the implicit method, the encoder directly processes the input regardless of modality. In particular, given a query VRT image x, we proceed as follows.

For the two-stage method:

(i) *Modality Translation*: Translate the query VRT image x into its corresponding CXR representation, denoted as $\mathsf{T}_{xy}(x)$.

(ii) *Joint Embedding Generation*: Map the image pair $(x, \mathsf{T}_{xy}(x))$ to a joint latent embedding, i.e., $z = \mathsf{NBP} = \mathsf{F}(x_i, \mathsf{T}_{xy}(x_i))$.

For the one-stage method:

(i) *Embedding Generation*: Map the query VRT image x to a latent embedding, i.e., $z = \mathsf{NBP} = \mathsf{E}(x, t)$, where t is the same as the constructed NBP Bank $\mathcal{Z}$.

Finally, we can identify the individual associated with x by performing a nearest neighbor search in the reference NBP bank $\mathcal{Z}$. This involves determining the NBP $\forall z_i \in \mathcal{Z}$ with the minimum distance d to z, i.e.

$$\mathrm{ID}(x) = \arg\min_i d(z, z_i). \tag{6}$$

4.4. Unified perspective of obtaining NBP via GCF

To consolidate the two fusion strategies introduced above, we provide a concise information-theoretic view [35] that unifies both implementations under the same conceptual objective. Recall that our goal is to estimate a modality-invariant identity representation, NBP, from paired skeletal input VRT x and soft-tissue input CXR y. Ideally, NBP should (i) retain identity information shared across both modalities, while (ii) suppressing modality-specific cues.

It can be expressed as following:

$$\max_{\mathsf{NBP}} I(\mathsf{NBP}; x, y) - \alpha\Big(I(\mathsf{NBP}; x \mid y) + I(\mathsf{NBP}; y \mid x)\Big). \tag{7}$$

where $I(\mathsf{NBP}; x, y)$ measures how much identity-relevant information the fused representation retains from both modalities, while $I(\mathsf{NBP}; x \mid y)$ and $I(\mathsf{NBP}; y \mid x)$ quantify modality-specific residual information that remains in NBP when conditioning on the other modality. Maximizing expression (7) therefore promotes a fused representation that emphasizes shared identity cues while minimizing modality-dependent bias.

In practice, we instantiate the above principle with two tractable components:

- A generative consistency term $\mathcal{L}_{\text{gen}}$, implemented either through explicit generation and reconstruction in the two-stage pipeline or diffusion denoising in the one-stage approach, ensuring that NBP captures information explainable by both inputs.

- A contrastive alignment term $\mathcal{L}_{\text{con}}$, which aligns NBP across modalities by maximizing a contrastive lower bound on the shared information.

The resulting composite training objective is

$$\mathcal{L} = \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{mod}}\mathcal{R}_{\text{mod}}, \tag{8}$$

where $\mathcal{R}_{\text{mod}}$ suppresses modality-specific residual signals that may leak into NBP. With this formulation, inference of the fused identity representation becomes

$$\text{NBP}^{\star} = \arg\max_{\text{NBP}} \left[\mathcal{L}_{\text{gen}}(\text{NBP} \mid x, y) + \mathcal{L}_{\text{con}}(\text{NBP}; x, y)\right]. \tag{9}$$

Importantly, the two-stage explicit fusion network $\mathsf{F}(\mathsf{T}, \cdot)$ and the one-stage diffusion encoder $\mathsf{E}(\cdot)$ are simply two parameterizations that instantiate the same underlying objective as expression (7). The former realizes $\mathcal{L}_{\text{gen}}$ through explicit translation and reconstruction, whereas the latter implements it implicitly via the denoising process within a generative diffusion model. Despite their architectural differences, both pathways optimize the same information-theoretic principle and therefore produce consistent NBP representations. This unified formulation also suggests future extensions—for instance, explicitly modeling aleatoric uncertainty [36] could further improve robustness when dealing with degraded or partially occluded skeletal inputs.

5. Evaluation

5.1. Datasets setup

5.1.1. Dataset

All VRT-CXR image pairs used in this work are real clinical data collected from West China Hospital of Sichuan University. The dataset consists of VRT-CXR image pairs collected from 1315 healthy individuals without pulmonary diseases or thoracic skeletal lesions. We randomly split it into training and test sets. The training set contains VRT-CXR image pairs from 1052 individuals, representing 80%. The remaining 263 VRT-CXR image pairs, representing 20% of the total, constitute the test set. In addition, we employ a large-scale CXR dataset provided by the National Institutes of Health Clinical Center, ChestX-ray8 [9], as a distractor database. As VRT-CXR image pairs are from healthy people, we only select CXRs with the "No Finding" label (without any apparent abnormalities) from ChestX-ray8 for distraction.

This study was performed with the approval of the ethics committee of the West China Hospital of Sichuan University. The requirement for informed consent was waived since this study was retrospective and did not involve the direct participation of patients.

5.1.2. Preliminary

Since this is a new task, we annotate some of our data and obtain the localization of thoracic skeletons in original VRT images and CXRs benefiting from the YOLOv8 [37]. The region of interest (ROI) criteria: The area surrounded by the horizontal tangent line of the upper edge of the first thoracic vertebra, the horizontal tangent line of the lower edge of the 12th thoracic vertebra, and the horizontal tangent line of the most lateral edge of the left and right ribs.

5.1.3. Data augmentation

To enhance the generalization and robustness of the model, we preprocess all of the input images for data augmentation. All images are resized to 256×256 with the bilinear interpolation. Color jittering involves changing the contrast of the image with a parameter set to 1.8. The probability of color inversion was 0.2. The image is flipped horizontally with a random probability of 0.5. To simulate the different rotation postures of the chest, it is randomly rotated $(-20, +20)$ degrees with a random probability of 0.3. In addition, VRT and translated VRT images are normalized with a mean of 0.3817 and a standard deviation of 0.3180; real and translated CXRs are normalized with a mean of 0.6425 and a standard deviation of 0.1613.

5.2. Metrics

Mean Average Precision (mAP) is a widely used metric in person re-identification when multiple gallery samples per identity exist. However, in our dataset, each identity is represented by a single CXR image in the gallery, making mAP degenerate (AP becomes binary). Therefore, we adopt Rank-k accuracy and Percentile Ranking Rate (PRR) to comprehensively evaluate retrieval performance [8].

5.3. Qualitative analysis

The VRT images and CXRs in the test set are organized as queries and the NBP bank, respectively. Then CXRs in ChestX-ray8 are gradually introduced into the bank to evaluate the robustness as distractors.

5.3.1. Alternative comparison

Given that this is the first comprehensive work on the thoracic skeleton identification task, there are no comparable specialized networks available. Therefore, we employ several classical identification models as baselines.

To begin with, since this is a person identification task that can be considered a biometric classification, much like face classification, directly adopting existing face recognition methods is a straightforward solution. In this manner, we compare our approach with some classical face identification frameworks, as shown in Fig. 8(a). Specifically, we employ various models as feature extractors and utilize a classical face classification loss, ArcFace [2], for comparison.

Another straightforward idea is to view this task as a cross-modal representation task, which could generally employ models based on contrastive triplet architecture with ResNet-18 as the backbone. To be specific, as shown in Fig. 8(b), we adopt two encoders to extract fine-grained modality-specific skeletal representations and use contrastive learning loss to optimize the representation process. Here, the VRT image serves as the anchor, the CXR of the same individual acts as the positive sample, and the CXR of another individual serves as the negative sample.

Considering these baselines only use real samples while CMT results may achieve better performance, as illustrated in Figs. 9(a)–9(b), we introduce the CMT results and extract features from each image for identifiable representations. The real CXR (VRT) image acts as the anchor, the CXR (VRT) image translated from the real VRT (CXR) of the same person acts as the positive sample, and the others' real CXR (VRT) images serve as the negatives.

These baselines cover appearance-based biometric identification, cross-modality contrastive learning, and generative augmentation strategies. Together, they represent all feasible fusion-free or weak-fusion paradigms applicable to this new task.

To validate the efficacy of both fusion strategies, we compare them against fusion-free baselines (CXR-only or VRT-only) in Table 1. Both explicit and implicit fusion substantially outperform their non-fused counterparts, confirming that cross-modality integration is essential for robust identification. The implicit variant achieves competitive performance with greater simplicity, highlighting its potential for resource-constrained forensic settings.

Explicit fusion outperforms implicit fusion due to several factors. The translation module directly reduces the severe domain gap by generating synthetic images that mimic the target modality [43], providing the fusion network with well-aligned input pairs. The pixel-level reconstruction loss further preserves fine-grained anatomical details critical for distinguishing individuals with similar thoracic skeletons. In contrast, the implicit method integrates modality alignment and identity preservation within a single diffusion process. While conceptually elegant, the diffusion model is primarily optimized for denoising rather than extracting discriminative features, and the contrastive constraint alone may not fully retain subtle morphological cues. Nevertheless, the implicit approach offers a simpler end-to-end alternative that still

Table 1
The performance of Ours and other alternatives on learning neural features of bones for identification. The queries are VRT images, and the matching pool is NBP bank constructed from the CXR database. A higher value is better.

Distractors	Rank-k Rate (%) ↑						
	$\|NBP - Bank\| = 263$ (+0)			+1k		+10k	
Methods	$k = 1$	$k = 10$	$k = 50$	$k = 10$	$k = 50$	$k = 10$	$k = 50$
Triplet [38]	0.38	4.94	27.76	4.18	20.53	3.80	17.11
IResNet-18 [39]	3.04	8.75	36.50	2.66	9.13	1.90	4.18
IResNet-50 [39]	2.28	10.27	29.66	3.42	12.17	1.14	3.80
LbA [40]	0.76	4.94	24.33	4.18	23.57	4.18	24.33
MobileFaceNet [41]	3.42	19.39	44.49	10.65	30.04	7.22	14.07
DiVE [42]	2.28	7.98	39.54	4.56	14.45	1.52	4.56
Niu et al. [8]	13.69	54.75	84.79	27.76	55.13	13.31	29.28
Implicit (Ours)	10.27	44.87	86.69	27.76	52.85	13.31	28.52
Explicit (Ours)	21.29	64.64	90.49	36.50	63.88	21.67	39.92

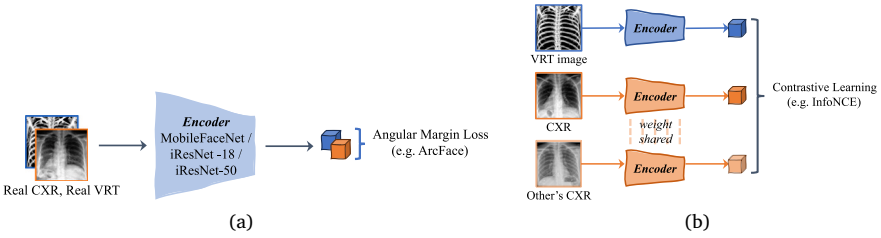

Fig. 8. (a) Leveraging the angular margin loss inspired by face classification. (b) Directly leveraging VRT images and CXRs for identification.

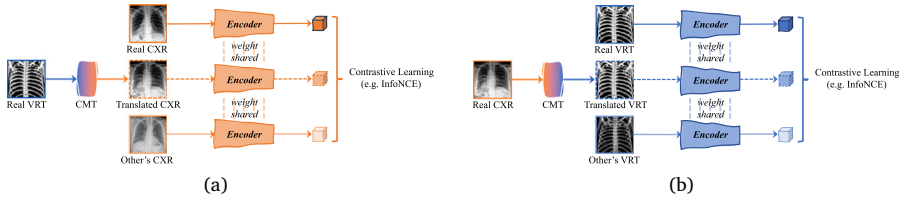

Fig. 9. (a) Solely identify in the CXR modality based on the translated CXR and the real ones. (b) Solely identify in the VRT modality leveraging the translated VRT images from the real CXRs.

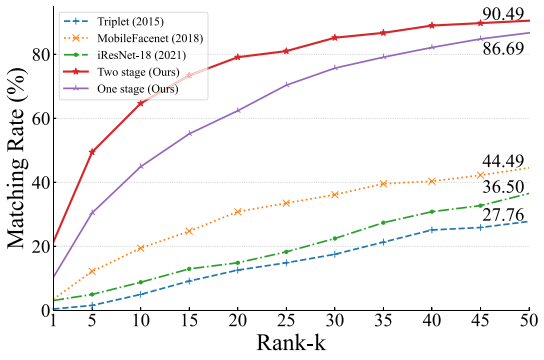

Fig. 10. Cumulative match characteristic curve.

outperforms all fusion-free baselines, improving Rank-50 accuracy from 84.79% [8] to 86.69%.

Additional experiments further contextualize these results. Comparisons between IResNet-18 and IResNet-50 [39] show that deeper networks do not necessarily yield significant performance gains, suggesting that model complexity is not the primary driver of accuracy in this task. We also compare our method with LbA [40] and DiVE [42], classic cross-modality person re-identification approaches.

5.3.2. Cumulative match characteristic curve

We select some with the best performance from all methods and plot the cumulative match characteristic (CMC) curve, as shown in Fig. 10. It illustrates the retrieval performance across increasing ranks. Our two-stage method consistently outperforms all baselines, achieving 90.49% at Rank-50. Notably, the one-stage method also maintains a high matching rate, indicating robust identity preservation through implicit fusion.

5.3.3. Ablation studies

We present an ablation study on the CMT module in Table 2. The results show that using CMT with only one modality (CXR or VRT) yields limited gains under large distractor sets, whereas the full two-modality injection achieves substantially higher robustness.

We also conduct the ablation study of the positive sample strategies for contrastive learning during the NBP extraction step, as shown in Table 3. Both present significant effectiveness of the CMT module. The augmented one uses the SimCLR [33] strategy without introducing cross-modality translation results. Additionally, we explore different weights for the contrastive and reconstructive loss, as shown in Table 4.

5.3.4. Related work comparison

We compared ours with some studies [16,17] based on CT images and CXRs for person identification in Table 5. Due to missing implementation details, our re-implementation exhibits minor deviations from the published numbers. To ensure fairness, we additionally report

Table 2
Ablation study on the impact of the CMT module under varying numbers of distractors. Results show Rank-50 identification rates (%) with default $|\mathsf{NBP} - Bank| = 263$.

Methods	+0	+1k	+5k	+10k
Triplet [38]	27.76	20.53	17.87	17.11
Triplet [38] + CMT (VRT)	62.36	27.00	9.51	6.46
Triplet [38] + CMT (CXR)	52.85	11.79	3.80	1.52
MobileFaceNet [41]	44.49	30.04	18.63	14.07
MobileFaceNet [41] + CMT	68.06	40.68	26.24	20.91
Explicit, CMF + CMT (Ours)	90.49	63.88	46.39	39.92

Table 3
Ablation study on the strategy of determining the positive samples for contrastive learning in the CMF step of the two-stage method under varying numbers of distractors. The augmented one in the second line introduces the augmentation strategy of SimCLR. Results show Rank-50 identification rates (%) with default $|\mathsf{NBP} - Bank| = 263$.

Positive samples	+0	+1k	+5k	+10k
Real CXR, translated VRT	19.39	0.00	0.00	0.00
Real Augmented [33]	49.05	11.03	4.18	2.28
Translated CXR, translated VRT	90.49	63.88	46.39	39.92

Table 4
Ablation study on the weights for the reconstruction loss and the contrastive learning loss of the two-stage method under varying numbers of distractors. λ_{re} denotes the weight for the reconstruction loss and λ_{cl} denotes the weight for the contrastive learning loss. Results show Rank-50 identification rates (%) with default $|\mathsf{NBP} - Bank| = 263$.

Settings		+0	+1k	+5k	+10k
λ_{re}	λ_{cl}				
1	0	33.08	0.00	0.00	0.00
0	1	90.11	63.50	46.01	37.26
1	1	89.73	60.84	39.92	34.60
1	3	90.49	63.88	46.39	39.92
1	5	90.11	61.22	43.35	36.12

Table 5
Our method was experimented with in a larger searchable bank and demonstrated superior performance. A lower p with a higher accuracy is better.

Method	$\lvert Query\rvert$	$\lvert Bank\rvert$	PRR, $p\%$
[16] (Steady MFV + BoW)	27	27	$p = 37.04$, 63.00%
[17] (CLAHE + DFT + Euclidean)	27	27	$p = 55.56$, 74.07%
Implicit (Ours)	263	1263	$p = 10$, 72.62%
	263	5263	$p = 10$, 72.24%
	263	10 263	$p = 10$, 71.86%
Explicit (Ours)	263	1263	$p = 10$, 77.19%
	263	5263	$p = 10$, 76.05%
	263	10 263	$p = 10$, 76.43%

the original results as provided in the papers. It demonstrates that our approach achieves superior performance even when handling more queries and a larger searchable bank.

5.3.5. Partial missing skeleton simulation & age analysis

The skeleton may be partially missing in some scenarios. To simulate this condition and validate robustness, query images were divided into 16 × 16 patches and randomly masked at varying ratios to mimic partial bone loss, as detailed in Fig. 11(a) and Table 6. Besides, we analyzed the performance across gender and various age groups, with results presented in Table 7 and Fig. 11(b), respectively.

Table 6
Rank-**50** Rate (%) with various mask ratios.

Mask ratio	Explicit	Implicit
0	90.49	86.69
0.10	88.97	83.65
0.15	83.65	77.57

Table 7
Results of genders.

	Male	Female
Total Num.	116	147
Hit Num. (Implicit)	101	127
Hit Num. (Explicit)	104	134
Rank-**50** Rate (Implicit, %)	87.07	86.39
Rank-**50** Rate (Explicit, %)	89.66	91.16

5.4. Quantitative analysis

We visualize the feature map of the best-performing experiment in Fig. 12. This visualization focuses on the same content as highlighted in Fig. 2(a) and confirms the consistency with human empirical assumptions.

We then visualize the top 5 samples and the last 5 samples of the identification results after introducing distractors, as shown in Fig. 13. The top five matches exhibit high anatomical similarity to the query, while the last five show low similarity, confirming the discriminative power of our method. Despite the deformations of thoracic skeletons and complex overlaps between VRT images and CXRs, which pose significant challenges for human identification, our proposed approach effectively performs the identification. This visualization underscores the robustness and accuracy of our method in handling challenging cases that are difficult for human observers to discern.

Failure analysis. We analyze the failure cases from our experiments, particularly those where the correct identity was ranked beyond Rank-50, and identify two common failure patterns: (i) Image quality degradation, where soft tissue shadows severely interfere with skeletal boundaries; (ii) Extreme anatomical similarity, where the thoracic skeletons of different individuals exhibit high similarity in overall morphology and key curvatures, even challenging forensic experts. This analysis clarifies the current limitations of the method, and future work should introduce an image quality assessment module or mine deeper fine-grained features.

6. Discussion

We have demonstrated that generative contrastive learning provides an effective paradigm for extracting this cross-modality identity representation from skeletal data. Our findings open new avenues for forensic identification and contribute to the growing intersection of medical imaging and machine learning.

While our framework achieves strong performance, several promising directions remain for enhancing the quality and utility of NBPs. First, the fidelity of the extracted features is fundamentally tied to the quality of the input images. Beyond improving bone imaging techniques [44], integrating more sophisticated backbones for skeletal feature extraction could yield finer-grained representations. For instance, architectures designed to assemble perceptual body parts [45] or to model dynamic skeletal structures [46] could be adapted to better capture the subtle morphological cues present in static medical scans. Second, the fusion process itself could be enriched. While we currently fuse VRT and CXR, incorporating additional modalities like MRI or PET could provide complementary textural or metabolic information. Recent advances in multi-modal interactive attention [47, 48] and vision-language models [49,50] offer promising frameworks for learning more robust, semantically grounded joint representations.

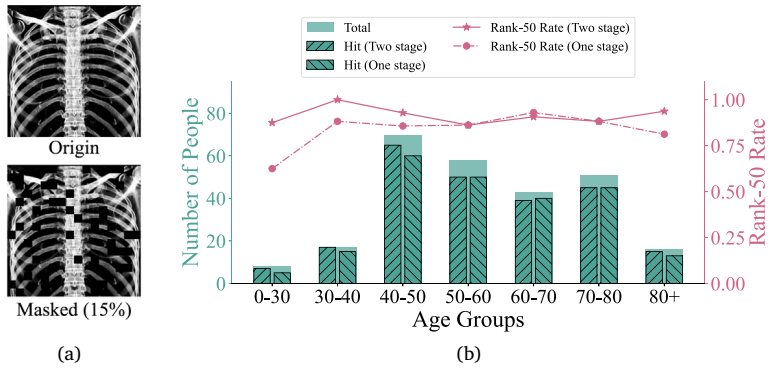

Fig. 11. (a) Masked VRT image. (b) Age fairness.

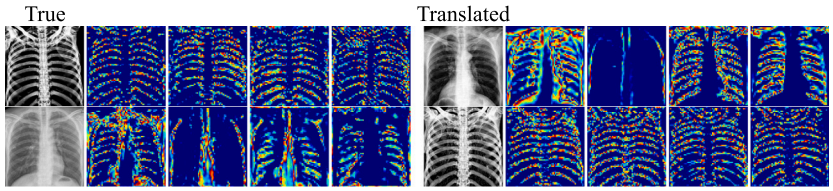

Fig. 12. Feature visualization.

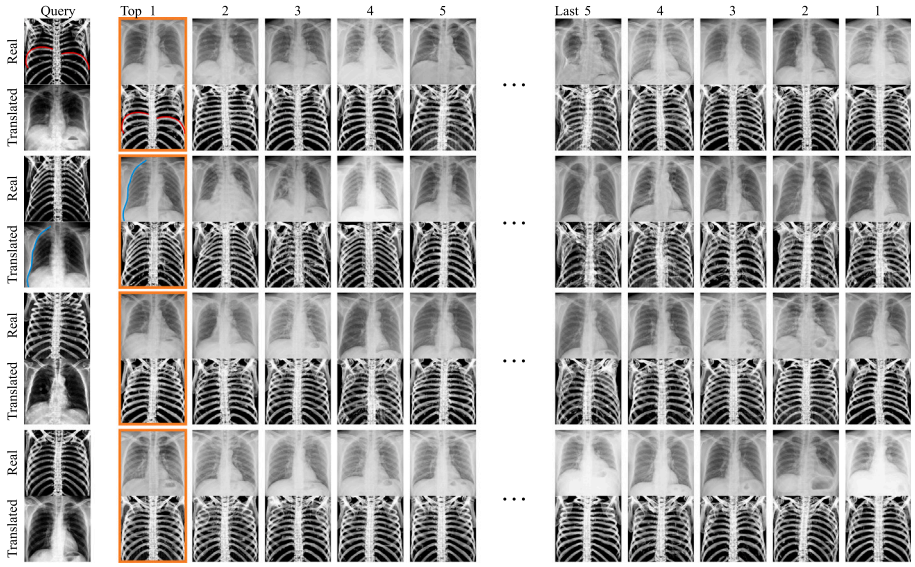

Fig. 13. Visualization of robust identification. The query pair consists of the real VRT image and its translated CXR. The candidate pairs consist of real CXRs with their translated VRT images. The orange rectangle represents the ground truth. The red and blue lines describe the rib boundary and the overall skeletal morphology, respectively. The top 5 identification results are similar to the query one, while the last are not.

Such approaches could help the model focus on key forensic landmarks, aligning its internal representations more closely with expert anatomical knowledge.

Beyond improving the core methodology, deploying such a system in real-world forensic contexts raises important practical considerations. Real cases often involve skeletal anomalies such as fractures or surgical implants. These irregularities can serve as powerful discriminative features, which our fine-grained model is well-suited to leverage. However, they also introduce challenges, such as increased intra-class variability due to bone healing over time. To this end, future work must evaluate model robustness on datasets with annotated anomalies and explore techniques like temporal modeling [46] to handle pre- and post-operative states. For time-sensitive scenarios like disaster response, computational efficiency is paramount. Our framework is inherently efficient at inference, requiring only the lightweight NBP encoder; the heavy generative components are discarded after training. To further enable edge deployment, we plan to explore model compression and lightweight architecture design. At the database level, our method is not limited to the single-sample-per-identity setup used in this study. It can naturally accommodate multiple gallery samples per person by either storing independent entries for voting or aggregating [51] NBPs into a robust prototype, enhancing its utility in real-world forensic databases. Ultimately, as we advance these technical capabilities, stringent privacy protection [52] and ethical oversight remain paramount, ensuring that any application upholds the rights and interests of individuals.

CRediT authorship contribution statement

Chaoqun Niu: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Dongdong Chen:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Jizhe Zhou:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jian Wang:** Writing – review & editing. **Quan-Hui Liu:** Writing – review & editing, Supervision. **Caiyang Yu:** Writing – review & editing. **Yuan Li:** Data curation. **Wei Ju:** Writing – review & editing. **Jiancheng Lv:** Writing – review & editing, Supervision, Resources, Funding acquisition.

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 was supported by the National Major Scientific Instruments and Equipments Development Project of National Natural Science Foundation of China under Grant 62427820.

Data availability

The code is available at *Neural-Boneprint*.

References

[1] R.S. Turingan, J. Brown, L. Kaplun, J. Smith, J. Watson, D.A. Boyd, D.W. Steadman, R.F. Selden, Identification of human remains using Rapid DNA analysis, Int. J. Legal Med. 134 (2020) 863–872.
[2] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[3] H.İ. Öztürk, B. Selbes, Y. Artan, Minnet: Minutia patch embedding network for automated latent fingerprint recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1627–1635.
[4] M. Fondevila, C. Phillips, N. Naverán, M. Cerezo, A. Rodríguez, R. Calvo, L. Fernandez, Á. Carracedo, M. Lareu, Challenging DNA: assessment of a range of genotyping approaches for highly degraded forensic samples, Forensic Sci. Int.: Genet. Suppl. Ser. 1 (1) (2008) 26–28.
[5] G.M. Hatch, F. Dedouit, A.M. Christensen, M.J. Thali, T.D. Ruder, RADid: a pictorial review of radiologic identification using postmortem CT, J. Forensic Radiol. Imaging 2 (2) (2014) 52–59.
[6] H. Fujimoto, Dental radiographic identification using ante-mortem CT, cone-beam CT, and MRI head and neck assessments, Forensic Imaging 26 (2021) 200465.
[7] J. Wen, F. Qin, J. Du, M. Fang, X. Wei, C.P. Chen, P. Li, MsgFusion: Medical semantic guided two-branch network for multimodal brain image fusion, IEEE Trans. Multimed. 26 (2023) 944–957.
[8] C. Niu, D. Chen, J. Zhou, J. Wang, X. Luo, Q.-H. Liu, Y. Li, J. Lv, Neural boneprint: Person identification from bones using generative contrastive deep learning, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7609–7618.
[9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.
[10] K. Packhäuser, S. Gündel, N. Münster, C. Syben, V. Christlein, A. Maier, Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data, Sci. Rep. 12 (1) (2022) 14851.
[11] Y. Ueda, J. Morishita, Patient identification based on deep metric learning for preventing human errors in follow-up X-ray examinations, J. Digit. Imaging 36 (5) (2023) 1941–1953.
[12] Z. Ali, N. Mourtzinos, B.B. Ali, D.R. Fowler, A pilot study comparing postmortem and antemortem CT for the identification of unknowns: Could a forensic pathologist do it? J. Forensic Sci. 65 (2) (2020) 492–499.
[13] N. Thampan, R. Janani, R. Ramya, R. Bharanidharan, A.R. Kumar, K. Rajkumar, Antemortem dental records versus individual identification, J. Forensic Dent. Sci. 10 (3) (2018) 158.
[14] A.H. Ross, A.K. Lanfear, A.B. Maxwell, Establishing standards for side-by-side radiographic comparisons, Am. J. Forensic Med. Pathol. 37 (2) (2016) 86–94.
[15] N. Shinkawa, T. Hirai, R. Nishii, N. Yukawa, Usefulness of 2D fusion of post-mortem CT and antemortem chest radiography studies for human identification, Jpn. J. Radiol. 35 (2017) 303–309.
[16] R. Ishigami, T.T. Zin, N. Shinkawa, R. Nishii, Human identification using X-Ray image matching, in: Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol. 1, 2017.
[17] H. Cho, T.T. Zin, N. Shinkawa, R. Nishii, Post-mortem human identification using chest x-ray and ct scan images, Int. J. Biomed. Soft Comput. Hum. Sci.: Off. J. Biomed. Fuzzy Syst. Assoc. 23 (2) (2018) 51–57.
[18] M.S. Macpherson, C.E. Hutchinson, C. Horst, V. Goh, G. Montana, Patient reidentification from chest radiographs: an interpretable deep metric learning approach and its applications, Radiol.: Artif. Intell. 5 (6) (2023) e230019.
[19] Y. Li, J. Wang, W. Liang, H. Xue, Z. He, J. Lv, L. Zhang, CR-GAN: Automatic craniofacial reconstruction for personal identification, Pattern Recognit. 124 (2022) 108400.
[20] E. Verna, M.-D. Piercecchi-Marti, K. Chaumoitre, C. Bartoli, G. Leonetti, P. Adalian, Discrete traits of the sternum and ribs: a useful contribution to identification in forensic anthropology and medicine, J. Forensic Sci. 58 (3) (2013) 571–577.
[21] N. Garoufi, A. Bertsatos, M.-E. Chovalopoulou, C. Villa, Forensic sex estimation using the vertebrae: an evaluation on two European populations, Int. J. Legal Med. 134 (2020) 2307–2318.
[22] L.C. Ebert, S. Franckenberg, T. Sieberth, W. Schweitzer, M. Thali, J. Ford, S. Decker, A review of visualization techniques of post-mortem computed tomography data for forensic death investigations, Int. J. Legal Med. 135 (5) (2021) 1855–1867.
[23] A. Neroladaki, D. Botsikas, S. Boudabbous, C.D. Becker, X. Montet, Computed tomography of the chest with model-based iterative reconstruction using a radiation exposure similar to chest X-ray examination: preliminary observations, Eur. Radiol. 23 (2013) 360–366.
[24] United Nations Scientific Committee on the Effects of Atomic Radiation, Sources and Effects of Ionizing Radiation, United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008 Report, Volume I: Report to the General Assembly, with Scientific Annexes a and B-Sources, United Nations, 2010.
[25] E. Çallı, E. Sogancioglu, B. van Ginneken, K.G. van Leeuwen, K. Murphy, Deep learning for chest X-ray analysis: A survey, Med. Image Anal. 72 (2021) 102125.
[26] F.A. Mettler, M. Bhargavan, K. Faulkner, D.B. Gilley, J.E. Gray, G.S. Ibbott, J.A. Lipoti, M. Mahesh, J.L. McCrohan, M.G. Stabin, B.R. Thomadsen, T.T. Yoshizumi, Radiologic and nuclear medicine studies in the United States and worldwide: Frequency, radiation dose, and comparison with other radiation sources—1950–2007, Radiology 253 (2) (2009) 520–531.
[27] C. Niu, Y. Li, J. Wang, J. Zhou, T. Xiong, D. Yu, H. Guo, L. Zhang, W. Liang, J. Lv, Multi-view adaptive bone activation from chest X-Ray with conditional adversarial nets, in: International Conference on Multimedia Modeling, Springer, 2023, pp. 399–410.
[28] G. Qin, H. Liu, W. Li, H. Zhang, Y. Guo, A virtual domain collaborative learning framework for semi-supervised microscopic hyperspectral image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 24–34.
[29] Y. Liu, C. Liu, J. Wen, L. Shen, B. Zhang, Y. Xu, Learning compact semantic information and reliable pseudo-labels for incomplete multi-view multi-label classification, IEEE Trans. Pattern Anal. Mach. Intell. (2026).
[30] J. Wen, C. Liu, S. Deng, Y. Liu, L. Fei, K. Yan, Y. Xu, Deep double incomplete multi-view multi-label learning with incomplete labels and missing views, IEEE Trans. Neural Netw. Learn. Syst. 35 (8) (2023) 11396–11408.
[31] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[32] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[33] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[34] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[35] N. Tishby, N. Zaslavsky, Deep learning and the information bottleneck principle, in: 2015 Ieee Information Theory Workshop (Itw), Ieee, 2015, pp. 1–5.
[36] C. Tang, T. Shen, X. Gong, C. Zhao, T. Zhang, DFMU: Distribution-based framework for modeling aleatoric uncertainty in multimodal sentiment analysis, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, 2025, pp. 8250–8258.
[37] G. Jocher, J. Qiu, A. Chaurasia, Ultralytics YOLO, 2023, URL https://github.com/ultralytics/ultralytics.
[38] E. Hoffer, N. Ailon, Deep metric learning using triplet network, in: Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3, Springer, 2015, pp. 84–92.

[39] I.C. Duta, L. Liu, F. Zhu, L. Shao, Improved residual networks for image and video recognition, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 9415–9422.
[40] H. Park, S. Lee, J. Lee, B. Ham, Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12046–12055.
[41] S. Chen, Y. Liu, X. Gao, Z. Han, Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices, in: Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13, Springer, 2018, pp. 428–438.
[42] W. Dai, L. Lu, Z. Li, Diffusion-based synthetic data generation for visible-infrared person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, (11) 2025, pp. 11185–11193.
[43] X. Wang, S. Zhang, H. Zhang, R. Wang, M. Li, C. Zhou, Q. Zhao, J.-Z. Zhou, Dehallu3D: Hallucination-mitigated 3D generation from single image via cyclic view consistency refinement, 2026, arXiv preprint arXiv:2603.01601.
[44] D. Chen, J. Tachella, M.E. Davies, Equivariant imaging: learning beyond the range space, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 4379–4388.
[45] R. Zhang, D. Lin, X. Wang, G. Baciu, C.P. Chen, P. Li, Accurate-PGNet: Learning to assemble perceptual body parts for accurate human skeleton establishment, IEEE Trans. Multimed. 27 (2024) 1706–1721.
[46] R. Zhang, D. Lin, X. Wang, R. Liu, B. Sheng, G. Baciu, C.P. Chen, P. Li, Temporal-interim pose synthesis and distillation for dynamic human pose estimation, IEEE Trans. Neural Netw. Learn. Syst. (2025).
[47] S. Li, T. Zhang, B. Chen, C.P. Chen, MIA-net: Multi-modal interactive attention network for multi-modal affective analysis, IEEE Trans. Affect. Comput. 14 (4) (2023) 2796–2809.
[48] K. Yang, Y. Luo, Z. Zhang, C.P. Chen, T. Zhang, Multimodal affect perception with large language model enhancement network, IEEE Trans. Affect. Comput. (2025).
[49] Z. Zhao, L. Deng, H. Bai, Y. Cui, Z. Zhang, Y. Zhang, H. Qin, D. Chen, J. Zhang, P. Wang, et al., Image fusion via vision-language model, in: International Conference on Machine Learning, PMLR, 2024, pp. 60749–60765.
[50] K. Lan, C.P. Chen, Z. Zhang, T. Zhang, Contrastive adversarial tuning: Enhancing discriminability and robustness of LLMs for emotion recognition in conversation, Pattern Recognit. (2026) 113025.
[51] S. Deng, J. Wen, C. Liu, K. Yan, G. Xu, Y. Xu, Projective incomplete multi-view clustering, IEEE Trans. Neural Netw. Learn. Syst. 35 (8) (2023) 10539–10551.
[52] X. Du, J. Zhu, J. Zhou, C.-m. Pun, Z. Lin, C. Wu, Z. Chen, J. Luo, Dp-trae: A dual-phase merging transferable reversible adversarial example for image privacy protection, IEEE Trans. Depend. Secur. Comput. (2025).