Concealed Face Analysis and Facial Reconstruction via Multi-Task Approach and Cross-Modal Distillation in Terahertz Imaging

Noam Bergman , Ihsan Ozan Yildirim , Asaf Behzat Sahin , Hakan Altan , Yitzhak Yitzhaky *

Posted Date: 26 January 2026

doi: 10.20944/preprints202601.1893.v1

Keywords: terahertz imaging; facial biometrics; multi-task learning; knowledge distillation; cross-modal fusion; deep learning; THz facial reconstruction

 doi:10.20944/preprints202601.1893.v1

Article

Concealed Face Analysis and Facial Reconstruction via Multi-Task Approach and Cross-Modal Distillation in Terahertz Imaging

Noam Bergman [1], Ihsan Ozan Yildirim [2], Asaf Behzat Sahin [3], Hakan Altan [4] and Yitzhak Yitzhaky [1,*]

[1] Department of Electro-Optical Engineering, School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, 1 Ben-Gurion Blvd, 8410501 Beer Sheva, Israel

[2] SDT Space and Defence Tech. Inc., SATGEB-2 Titanyum C Blok İhsan Dogramaci Bulv No 37, Ankara, Turkey

[3] Department of Electrical Electronics Engineering, Ankara Yıldırım Beyazıt University, Ankara, Turkey

[4] Department of Physics, Middle East Technical University, Ankara, Turkey

* Correspondence: ytshak@bgu.ac.il

Abstract

Terahertz (THz) sub-millimeter wave imaging offers unique capabilities for stand-off biometrics through concealment, yet they suffer from severe sparsity, low resolution, and high noise. To address these limitations, we introduce a novel unified Multi-Task Learning (MTL) network centered on a custom shared U-Net-like THz data encoder. This network is designed to simultaneously solve three distinct critical tasks on concealed THz facial data, given a limited dataset of approximately 1,400 THz facial images of 20 different identities. The tasks include concealed face verification, facial posture classification, and a generative reconstruction of unconcealed faces from concealed ones. While providing highly successful MTL results as a standalone solution on the very challenging dataset, we further studied the expansion of this architecture via a cross-modal teacher-student approach. During training, a privileged visible-spectrum teacher fuses limited visible features with THz data to guide the THz-only student. This distillation process yields a student network that relies solely on THz inputs at inference. The cross-modal trained student achieves better latent space in terms of inter-class separability compared to the single-modality baseline, but with reduced intra-class compactness, while maintaining a similar success in the task performances. Both THz-only and distilled models preserve high unconcealed face generative fidelity. The implementation code and trained weights are provided at https://github.com/noamberg/thz-face-distil.

Keywords: terahertz imaging; facial biometrics; multi-task learning; knowledge distillation; cross-modal fusion; deep learning; THz facial reconstruction

1. Introduction

Imaging in the sub-millimeter wave (Sub-MMW) or Terahertz (THz) frequency ranges has emerged as a powerful technology for security and surveillance applications. A key benefit of THz radiation is its non-ionizing nature, making it safe for screening people [1]. The characteristics of THz radiation, make it usable in various application in medical imaging [2] and industrial quality control [3]. Its primary advantage in security applications lies in the ability of THz waves to penetrate common clothing and other non-metallic, non-polar dielectric materials, revealing objects or features concealed underneath [4]. THz imaging has been demonstrated for biometric applications like facial recognition, even under occluding materials like masks, scarves, or balaclavas.

Prior research has confirmed that active imaging, particularly in the 340 GHz frequency band, can successfully penetrate common concealing fabrics to capture person-specific facial structures,

 doi:10.20944/preprints202601.1893.v1

thereby motivating the development of advanced biometric verification systems [5]. Early-stage approaches in this domain often relied on traditional computer vision techniques, such as template matching [6,7]. These methods, sometimes using metrics like the Structural Similarity Index (SSIM), operate by performing a pixel-wise or structural comparison of a concealed face image against a gallery of pre-recorded, unconcealed reference images. While this established a baseline capability, the performance of template matching is fundamentally limited [8]. These methods are highly sensitive to variations in a subject's posture, a challenge typically addressed in the visible domain by geometric frontalization techniques [9], yet they remain largely ineffective in the sparse Terahertz domain, where the lack of high-fidelity textural landmarks prevents standard alignment. Furthermore, they do not fully exploit the information encoded in the Sub-MMW signal, which contains more than just topographic data [10].

For example, recent work has demonstrated that this latent information includes complementary data from both the amplitude and phase of the multi-spectral THz signal; low-frequency phase can provide precise depth information while the corresponding amplitude can delineate finer object contours, with their fusion leading to improved image restoration [11]. Further expanding on the concept of latent biometric markers, another research has demonstrated that unique and discriminative information is also present in the radiometric signature of the human thorax, with multi linear analysis of this region in MMW images, showing more promising identification results than the facial region alone [12]. Complementing these holistic, texture-based methods, alternative feature-based approaches have also been explored, where a small set of discriminative, distance-based features is extracted from the geometry of the body silhouette, with measures like height and waist width proving effective for user separation [13]. This principle of analyzing unique signal responses has also been applied to fingerprint liveness detection, where the distinctive back-reflection from the inner layers of the epidermis, observable in the THz time-domain signal, provides a reliable way to distinguish real fingers from artificial spoofs [14].

1.1. Applying Deep Learning Approaches for Terahertz and Sub-Millimeter-Wave Imaging

Deep learning offers a promising alternative, with the potential to learn robust, discriminative features directly from the noisy, low-resolution data. Previous work has explored deep learning for specific, isolated tasks like face verification or posture estimation [10,15]. However, a more holistic approach is needed to fully realize the potential of Sub-MMW facial analysis for real-world applications. A single system that can simultaneously verify identity, determine the subject's posture or orientation, and reconstruct a visual likeness for human analysis would represent a significant leap forward.

Unlike template matching, deep learning models can automatically learn discriminative features directly from the raw or minimally processed THz images. This data-driven approach allows them to be inherently more robust to the variations that degrade the performance of hand-crafted feature-based methods [10]. Research has shown success in adapting pre-trained CNNs like AlexNet and VGG-face to extract meaningful features from millimeter-wave body and face images, a technique known as transfer learning [15]. More advanced, specialized architectures are also being developed, such as the Cross-Feature Fusion Transformer YOLO (CFT-YOLO), which is specifically designed to handle the unique challenges of THz images, including low resolution and high noise levels [16]. These modern frameworks enable not just more reliable verification but also more complex tasks, such as the pixel-wise reconstruction of a clear facial image from the occluded THz data [11]. Further innovation in this area is demonstrated by the mmID system, which employs a conditional generative adversarial network (cGAN) to reconstruct high-resolution images of the human body from low-resolution mmWave radar data, enabling more accurate human identification. This system first estimates the spatial spectrum using the Multiple Signal Classification (MUSIC) algorithm and then uses this as input for the cGAN to generate a high-resolution image, which is then used for identification with a CNN-based classifier [17].

Another critical application is in security, for the detection of concealed objects. The low quality

 doi:10.20944/preprints202601.1893.v1

of THz images, characterized by high noise and poor resolution, makes accurate detection challenging. To overcome this, research has shown the effectiveness of a two-step approach: first, preprocessing the THz images to enhance their quality, and then using a deep learning model for detection. One study successfully used Non-Local Mean (NLM) filtering to suppress noise in the images, followed by the YOLOv7 algorithm to detect concealed objects, achieving high recognition accuracy. This highlights the value of combining image processing with advanced detection models like YOLOv7 to create practical systems for security screening [18]. To combat the challenge of limited and low-quality training data, another approach involves generating synthetic active millimeter-wave imaging (AMWI) datasets using depth-based and distance-based simulation methods, which are then used to train a CNN for more accurate concealed object recognition [19]. Furthermore, to enhance real-time concealed object detection in passive terahertz security images, an improved Single Shot MultiBox Detector (SSD) has been proposed, which replaces the original VGGNet encoder with ResNet-50, incorporates a feature fusion module, and utilizes a hybrid attention mechanism and Focal Loss to achieve a mean average precision of 99.92% at 17 frames per second [20]. To address the challenge of the human body acting as background noise, another effective strategy involves a parallel method where an improved threshold segmentation technique first locates the human body, after which a Faster R-CNN model is used to detect concealed objects, significantly improving both detection speed and accuracy [21].

Shifting focus to image enhancement, a super-resolution reconstruction algorithm based on a multi-scale channel attention (MS-CA) residual network has been proposed to improve terahertz images for postal security inspection, achieving higher peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) compared to traditional deep-learning algorithms [22]. Another work in heterogeneous face recognition (HFR) has largely focused on direct feature matching between thermal and visible domains [23], often ignoring the structural guidance that high-fidelity modalities can provide. More recent advances in multi-task learning [24] suggest that auxiliary objectives, such as pose estimation, can regularize feature learning in data-scarce environments. However, the integration of cross-modal knowledge distillation [25] within a multi-task framework for sub-millimeter wave biometrics remains under-explored.

While prior research has demonstrated the potential of THz and Sub-MMW imaging for isolated tasks such as concealed object detection, image reconstruction, or biometric identification, these efforts have largely treated each problem separately. A greater potential for robust, real-world security surveillance lies in a more holistic approach that synthesizes these objectives. This paper introduces a unified Multi-Task Learning (MTL) framework specifically designed to address this challenge in the context of concealed facial analysis.

Our primary contributions are as follows: (i) Generating for the first time, reconstructed unconcealed faces from concealed ones (covered by Balaclava), and carrying facial posture classification in THz images, which are characterized by high noisiness and sparse geometry that obscure fine-grained facial biometrics (as seen in Figure 1). (ii) We propose a pioneering end-to-end deep learning model that simultaneously performs three critical tasks on concealed facial THz images: identity verification, facial posture classification, and unconcealed face reconstruction. (iii) We introduce a novel Cross-Modal Knowledge Distillation model for THz images, where a "privileged" visible-spectrum teacher guides the training of a THz student encoder. By distilling semantic knowledge from the high-fidelity visible domain into the sparse THz latent space, within an MTL framework, our model learns a rich, structure-aware representation of the sparse THz data employed to achieve the three objectives.

The proposed framework moves beyond the limitations of earlier template-matching classical methods and single-focus deep learning methods, presenting a combined novel approach for Sub-MMW facial analysis.

 doi:10.20944/preprints202601.1893.v1

Figure 1. Unconcealed THz images of subject ID #10 across five distinct facial postures, and their visual counterparts. The top two rows present THz *unconcealed* faces across all facial postures with their corresponding visible images below, while the bottom two rows present the THz *concealed* faces across all facial postures with their corresponding visible images below. The THz samples demonstrate the inherent challenges of the active 340 GHz modality, characterized by low spatial resolution, high acquisition noise, and sparse geometric representations that obscure fine-grained facial biometrics.

2. Materials and Methods

2.1. THz Image Acquisition System

The data used in this study was captured using an active frequency-modulated continuous-wave (FMCW) imaging system operating at a center frequency of 340 GHz [5]. The system was configured in a confocal Gregorian optical geometry, designed to acquire head-region intensity images at a nominal standoff distance of approximately 6 meters. The transmitter chain consists of Schottky diode multipliers, a pyramidal horn, and a 3 dB high-directivity coupler for transmit-receive coupling. The receiver performs subharmonic mixing to detect the reflected signal. The Sub-MMW beam is focused using a large elliptical reflector and a parabolic sub-reflector. Optical simulations in ZEMAX were used to validate the beam propagation, confirming a beam spot of approximately 1.5 cm in diameter at the focus, with a depth of focus of about 41 cm. This optical resolution is sufficient to preserve the macro-geometry of the human face across various facial postures.

To acquire an image, two large galvanometric mirrors mechanically steer the beam across a 45x45 cm field of view, scanning at 32 Hz (horizontal) and 1 Hz (vertical). This process yields frames with 64 vertical scan lines, which are then up-sampled to 256x64 pixels to maximize the information content. To improve the signal-to-noise ratio (SNR), three consecutive frames, each captured in 0.5 seconds, are averaged. The final reconstructed images have a dynamic range of approximately 8.3 dB for facial scans. Safety assessments confirmed that the system operates well below the ICNIRP

 doi:10.20944/preprints202601.1893.v1

exposure limits for safety, with a peak power density of ~2.0 mW/cm^2 and an effective exposure of ~2.5x10^{-4} s per cm^2 per frame.

2.2. Dataset

The dataset comprises approximately 1,400 total Sub-MMW images acquired from 20 different identities [5]. For each identity, images were captured under two conditions: unconcealed (no facial covering) and concealed (wearing a balaclava). To ensure robustness to facial posture variations, data was collected across five distinct face postures: front-left, front-center, front-right, full-left (profile), and full-right (profile). This resulted in approximately 35 unconcealed and 35 concealed images per identity, distributed across the five postures. Figure 1 presents an example of both concealed and unconcealed Visible and THz paring face images across five distinct facial postures. For our experiments, a stratified split was performed, holding out 15% of the data for testing. The test set included one concealed and one unconcealed image for each identity in each of the five postures. The remaining data was used for training and validation, with a balanced representation of identities and postures.

In addition to the Sub-MMW image data, our dataset includes also a smaller collection of visible-range images, comprising one concealed and one unconcealed sample for each identity and facial posture. This yields 10 visible-range images per identity, 2 for each of the five facial postures, and a total of 200 images overall. Integrating this auxiliary data enables us to extend the proposed approach into a multimodal framework, in which knowledge from the visible spectrum can be distilled into the THz domain using a teacher–student training paradigm, while preserving a THz-only configuration at inference time.

All raw images were cropped to the face Region of Interest (ROI). The resulting grayscale ROIs were then standardized to a uniform size of 64x64 pixels using bicubic interpolation [26]. This resolution was chosen as a trade-off between preserving facial details and managing the computational complexity of the deep learning model.

2.3. A Unified Multi-Task Learning Using Sub-MMW Imagery Only

To fully utilize the information embedded in Sub-MMW imagery, we developed a novel Multi-Task Learning (MTL) framework. This architecture trains a unified model to simultaneously master three interrelated objectives: concealed face verification, facial posture classification (across both concealed and unconcealed states), and the high-fidelity reconstruction of unconcealed faces from their concealed counterparts. By compelling the network to distill a shared feature representation relevant to all three tasks, our approach enhances data efficiency, improves generalization, and mitigates overfitting, a common challenge in sparse-data domains [24]. The architecture utilizes a shared Convolutional Neural Network (CNN) encoder that projects input Sub-MMW images into a high-dimensional latent space. This mutual representation feeds into three specialized heads: (1), a verification head that generates discriminative embeddings for identity matching, (2), a classification head that predicts discrete head poses, and (3), a reconstruction head that functions as a generative decoder which generates unconcealed facial images from concealed ones. This joint optimization strategy forces the encoder to disentangle identity-invariant features from pose-dependent variations, yielding a more robust and semantically rich understanding of facial data than is achieved through isolated single-task learning.

In our framework, each training instance is organized in a triplet formulation of anchor–positive–negative structure, that jointly encodes identity information, facial posture variation, and modality asymmetry between concealed and unconcealed facial observations. Specifically, the *anchor* corresponds to a concealed face image of a given identity under a particular facial posture. The *positive* is an unconcealed image of the same identity captured under the same posture. The *negative* is an unconcealed image drawn from a different identity but matched in posture. This structured sampling serves three complementary purposes that enable a unified multi-task learning formulation. First, the anchor–positive pairing provides explicit supervision for concealed-to-

 doi:10.20944/preprints202601.1893.v1

unconcealed face verification, requiring the model to extract identity-preserving features despite missing or partially occluded facial information. Second, because all samples are annotated with facial posture labels, the framework naturally supports posture classification, allowing the network to disentangle identity-relevant from pose-dependent cues. Third, the pairing of concealed anchors with their corresponding unconcealed positives provides a reconstruction target that enables unconcealed face synthesis conditioned on concealed inputs, improving both representation separation and robustness across modalities. By feeding all tasks through a shared encoder and by structuring each triplet to reflect identity, pose, and visibility factors, our formulation unifies the three learning objectives under a single sampling.

2.3.1. A Shared Feature Encoding via a Shared CNN Encoder

At the core of our MTL framework lies a shared U-Net—type CNN, which serves as the foundational feature encoder for all subsequent tasks, considered as decoders (Figure 3). This shared encoder transforms the input 64x64 pixel Sub-MMW facial image into a dense, informative latent representation through a hierarchical series of convolutional layers that progressively abstract low-level textures into high-level semantic features. This architectural strategy, widely validated in multi-task visual facial analysis [24], generates a compact, distilled feature vector that serves as a comprehensive basic descriptor of the essential image content. This latent vector is broadcast simultaneously to the verification, classification, and reconstruction heads (Figure 3, upper part), ensuring that all tasks rely on shared features, where insights from one task improve the others.

2.3.2. Concealed Face Verification Head

The verification head is tasked with authenticating identity by matching a concealed face (e.g., balaclava-masked) against an unconcealed reference (Figure 3, left column). Leveraging a 128-dimensional shared latent representation from the encoder, this module employs a linear projection followed by L2-normalization to constrain embeddings to a unit hypersphere. This normalization is critical for metric learning, ensuring that similarity is measured solely by angular distance (cosine similarity) rather than vector magnitude, thereby improving robustness to illumination and contrast variations [27].

To learn an identity-invariant representation across concealed and unconcealed facial observations, we adopt a supervised contrastive learning objective (SupCon) [28] that generalizes beyond the classical triplet loss. Each sample in the batch is indexed by $n \in 1, \dots, N$ and is associated with an identity label y_n. Let $\tilde{z_n}$ denote the normalized embedding corresponding to sample n. Posture and concealment attributes are treated as intra-identity variations and do not define the label space. For any chosen anchor sample with index n, we define a ***positive set*** that includes all other samples in the batch that share the same identity label as the anchor

$$P(n) = \{m \in 1, \dots, N \mid m \neq n, y_m = y_n\} \tag{1}$$

And a contrastive set that includes all remaining samples in the batch except the anchor

$$A(n) = \{m \in 1, \dots, N \mid m \neq n\} \tag{2}$$

Thus, $P(n)$ contains same-identity samples (both concealed and unconcealed views across postures), whereas $A(n)$ serves as the denominator of the contrastive SoftMax and includes all other samples, spanning both same-identity positives and different-identity negatives. SupCon applies its loss once per anchor index, systematically rotating through all samples in the batch. The supervised contrastive loss for the batch [28] is then expressed as

$$L_{SupCon} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{-1}{|P(i)|}\cdot\sum_{p\in P(i)} log\frac{exp\left(\text{sim}\left(\tilde{z_i},\tilde{z_p}\right)/\tau\right)}{\sum_{k\in A(i)} exp\left(\text{sim}\left(\tilde{z_i},\tilde{z_p}\right)/\tau\right)}\right] \tag{3}$$

where $\tilde{z_i}, \tilde{z_p}$ and $\tilde{z_k}$ are the normalized embeddings corresponding to same-identity samples $P(n)$ and different-identity samples $A(n)$. The outer summation over N iterates once over every sample

 doi:10.20944/preprints202601.1893.v1

in the batch, treating each sample in turn as an anchor. The set $P(n)$ contains all other samples that share the same identity label as the anchor n, and therefore defines the positive pairs used to pull same-identity representations together. The set $A(n)$ contains all remaining samples in the batch and forms all the contrastive pairs used in the denominator. These include both the positives and all different-identity negatives, which push the anchor away from unrelated identities. The inner logarithmic Softmax term compares the cosine similarity (sim) between the anchor and each positive to the similarities between the anchor to all samples in the contrastive pairs. The temperature parameter τ controls the sharpness of this Softmax, with smaller values increasing the emphasis on harder negatives. The normalization by $|P(n)|$ ensures that all anchors contribute equally regardless of how many positive views of the same identity appear in the batch. Together, these components encourage a consistent embedding space in which all concealed and unconcealed views belonging to the same identity are compactly clustered, while representations of different identities remain well separated.

2.3.3. Facial Posture Classification Head

This head is designed for the classification of a facial posture into one of five distinct categories. The architectural design centers on a 2D Convolutional layer followed by a linear layer that directly precedes a Softmax activation function. This stage of the network is responsible for transforming the high-dimensional feature vectors, which are learned by the preceding layers, into a probabilistic distribution across the five predefined postural classes.

To train this classification head, a Categorical Cross-Entropy (CCE) loss function, L_{CCE}, is employed. This loss function is common in multi-class classification tasks, as it quantifies the dissimilarity between the predicted probability distribution and the ground-truth distribution.

2.3.4. Unconcealed Face Reconstruction Decoder

The primary objective of this decoder is to synthesize the unconcealed facial image from latent representations derived from concealed Sub-MMW sensor data. The architecture functions as a generative decoder, taking a compressed feature vector from the encoder and progressively up-samples it through a sequence of convolutional layers to produce a 64x64 pixel resolution image. The reconstruction decoder utilizes a U-Net-like architecture, processing multi-scale encoder features via a bottleneck Residual Dense Block and progressively recovering spatial resolution through PixelShuffle upsampling layers interleaved with skip connections and convolutional fusion blocks, terminating in a sigmoid-activated reconstruction [29]. To enhance the reconstruction fidelity and preserve fine facial details, this decoder incorporates a Residual Dense Block (RDB) as the bottleneck module between the encoder and the decoder parts (right part , Figure 2), which utilizes densely connected convolutional layers where each layer receives feature maps from all preceding layers, enabling efficient feature reuse and gradient flow [29]. The RDB implements local feature fusion via 1×1 convolutions to stabilize training, followed by residual connections for hierarchical feature learning. This dense connectivity and residual architecture facilitate robust feature propagation throughout the decoder network, preserving fine-grained facial characteristics during the reconstruction process [38]. The composite loss, L_{rec}, is formulated as follows:

$$L_{rec} = \alpha \cdot L_{Charbonnier}(x, \hat{x}) + \beta \cdot \left(1 - \text{SSIM}(x, \hat{x})\right) + \gamma \cdot L_{LPIPS}(x, \hat{x}) \quad (4)$$

The reconstruction loss L_{rec} combines a robust Charbonnier term for pixel-wise fidelity, a (1–SSIM) term is the L_{SSIM} for preserving structural similarity, and the L_{LPIPS} term serves for perceptual consistency between x (ground-truth) and $\hat{x}$ (predicted) unconcealed THz faces, weighted by α, β, and γ, respectively. Together, these terms jointly enforce low-level accuracy, structural integrity, and high-level perceptual quality in the reconstructed images. The Charbonnier loss component enforces pixel-level fidelity between the reconstructed and ground-truth images. It is a robust variant of the L1 loss that is less sensitive to outliers, ensuring precise pixel-wise correspondence. Its explicit expression is

 doi:10.20944/preprints202601.1893.v1

$$L_{\text{Charbonnier}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(I_{Pred}^{(i)} - I_{\text{GT}}^{(i)}\right)^2 + \epsilon^2} \tag{5}$$

This calculates the average error over all pixels by summing the square-rooted, squared differences between each reconstructed pixel and its ground-truth counterpart, with a small constant ϵ to ensure differentiability and stability. The Structural Similarity Index Measure ($SSIM$) Loss component aims to preserve the local structural similarity between the images, ensuring that textures and fine details are realistically rendered. $SSIM$ evaluates luminance, contrast, and structure. The loss is formulated as $1 - SSIM$ to be minimized. The $SSIM$ index between two images, x and y, is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

Here, μ and σ represent the local means and standard deviations, σ_{xy} is the cross-covariance,C_1 and C_2 are small constants that stabilize the division. The Learned Perceptual Image Patch Similarity ($LPIPS$) Loss leverages a pre-trained deep neural network to measure the perceptual similarity between the reconstructed and ground-truth images. By comparing feature maps from various layers of a network trained on a large-scale image dataset, $LPIPS$ provides a metric that aligns more closely with human perception of image quality [30]. This promotes the generation of visually plausible and realistic results. The loss is computed as:

$$L_{LPIPS}(X, X_0) = \sum_l \frac{w_l \cdot 1}{H_l W_l} \sum_{h,w} \left| \hat{y_{h,w}^l} - \hat{y_{0,h,w}^l} \right|_2^2 \tag{7}$$

In this equation, y^l and y_0^l are the channel-normalized feature activations from layer l for the two images. The squared L2 distance is computed for each spatial location, averaged, and then combined as a weighted sum across layers, with weights w_l.

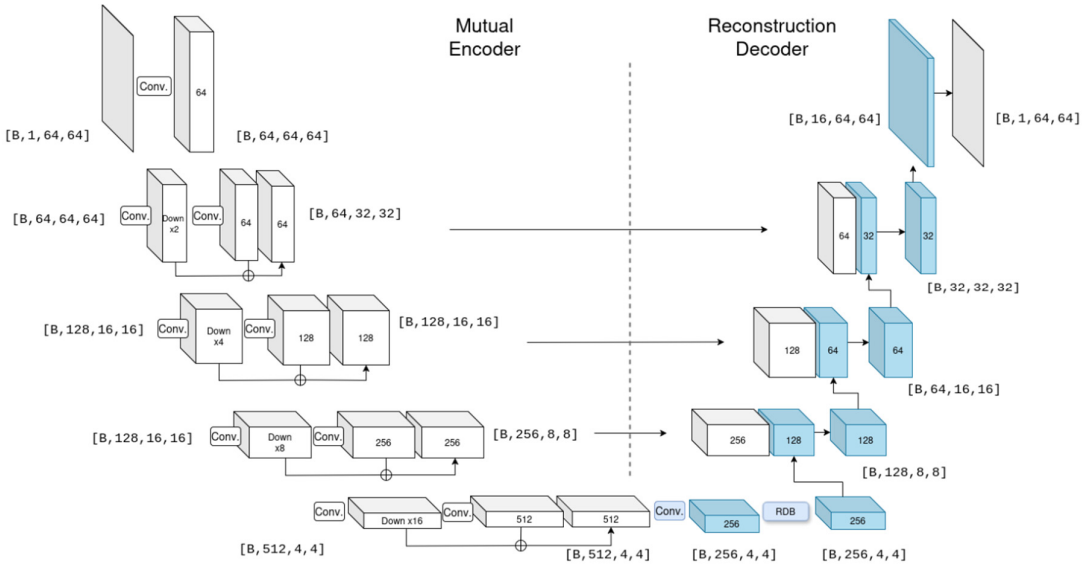


Figure 2. The architecture of the shared CNN encoder (left part, white boxes), combined with the unconcealed face reconstruction decoder (right part, blue boxes). This design is inspired by a U-Net structure, where the encoder uses convolutional layers for learnable down-sampling, and the decoder uses pixel-shuffle layers for up-sampling. A Residual Dense Block (RDB) is placed in the bottleneck for advanced feature extraction.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2026 doi:10.20944/preprints202601.1893.v1

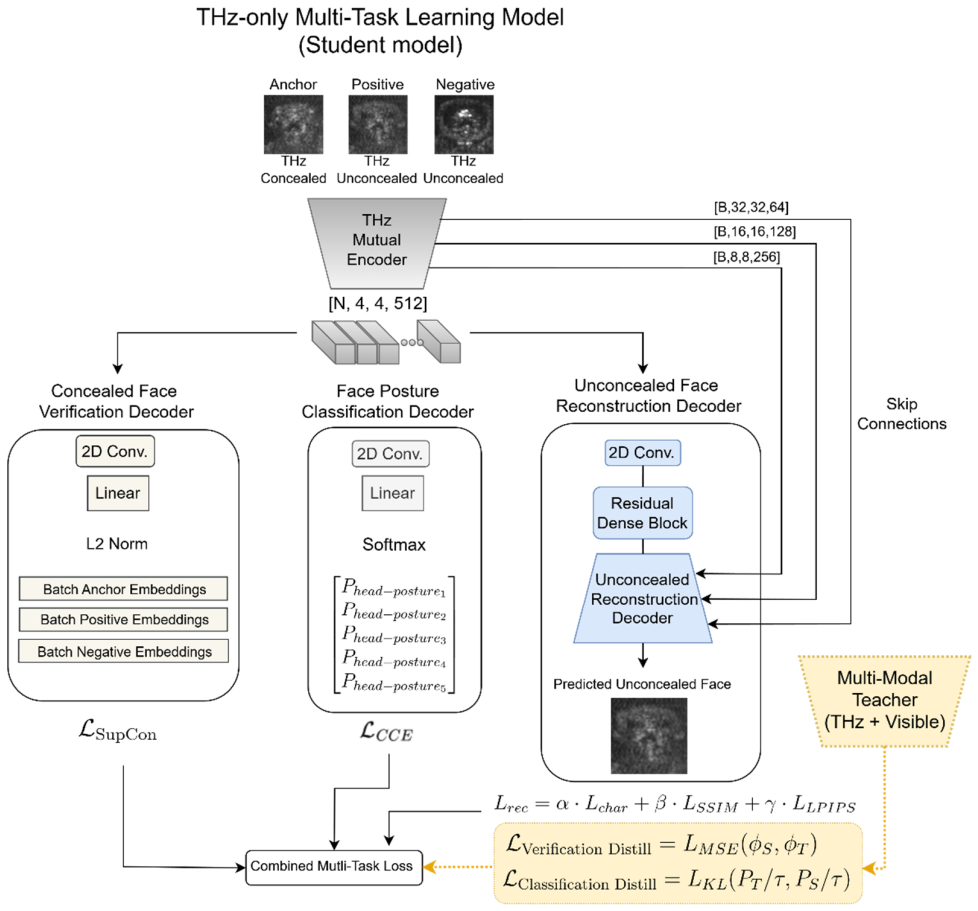


Figure 3. A schematic overview of the THz-only Multi-Task Learning architecture (also, the student model). The model processes a triplet of Sub-MMW facial images (Anchor [Concealed, subject i], Positive [Unconcealed, subject i], Negative [Unconcealed, subject j]) through a mutual CNN encoder, which extracts a unified latent representation. This shared feature vector is simultaneously projected into three task-specific branches. The mutual encoder ensures that feature learning is regularized by the competing yet synergistic objectives of all three tasks.

2.4. Integration of a Visible-Range Modality: A Teacher—Student model

Although the proposed multi-task model operates on THz Sub-MMW images only at inference, we introduce a visible–THz learning multi-modal framework that exploits visible-range facial images only during training (Figure 3). The goal is to examine how the THz encoder can benefit from the richer structural cues present in visible images, without requiring an additional sensor at deployment [31,32].

2.4.1. Dual-Encoder Architecture

Our framework adopts a dual-encoder design to enable effective cross-modal knowledge transfer during training [33]. The THz encoder E_{thz} operates on both concealed and unconcealed THz facial images and serves as the sole feature extractor at inference time (Figure 4). Complementing this, a visible-range encoder E_{vis} processes both masked and unmasked visible-spectrum images, but is utilized only during training to provide auxiliary supervisory signals. Each modality's encoder (Figure 4) produces a spatial feature representation that is subsequently projected into a sequence of latent tokens T_{thz}, T_{vis}, forming the basis for the multi-modal knowledge distillation process, a strategy akin to the cross-modal teacher-student frameworks proposed by [34].

 doi:10.20944/preprints202601.1893.v1

2.4.2. Cross-Modal Fusion via Symmetric Cross-Attention

To robustly transfer semantic structure from the visible domain into the THz feature space, we employ a Dual Cross-Attention mechanism that symmetrically fuses features guided by each modality [31]. This module consists of two parallel attention branches: a Visible-Guided branch where visible tokens act as queries to extract relevant details from the THz context, and a THz-Guided branch where THz tokens query the visible features to hallucinate missing high-frequency textures [32].

Let T_{vis} and T_{thz} denote the token sequences from the visible and THz encoders, respectively. The Visible-Guided cross-attention is computed as:

$$Q_{vis} = T_{vis}W_Q^{vis}, K_{thz} = T_{thz}W_K^{vis}, V_{thz} = T_{thz}W_V^{vis}$$
$$T_{vis\to thz} = Softmax\left(\frac{Q_{vis}K_{thz}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot V_{thz} \quad (8)$$

Where Q_{vis} Projects the visible-domain token features T_{vis} into a query embedding space using a learnable visible-query projection matrix W_Q^{vis} to produce visible queries. K_{thz} projects the THz-domain token features T_{thz} into a key embedding space using a projection matrix W_K^{vis} (shared with or aligned to the visible domain) to obtain cross-modal keys K_{thz}. V_{thz} Projects the THz-domain token features T_{thz} into a value embedding space via W_V^{vis}, producing cross-modal values V_{thz} that carry THz content information. $T_{vis\to thz}$ computes cross-attention from visible queries to THz keys, normalized by $\sqrt{d_k}$ and passed through a Softmax, then aggregates THz values V_{thz} to produce THz-aware visible features $T_{vis\to thz}$. Symmetrically, the THz-Guided cross-attention is defined as:

$$Q_{thz} = T_{thz}W_Q^{thz}, K_{vis} = T_{vis}W_K^{thz}, V_{vis} = T_{vis}W_V^{thz}$$
$$T_{thz\to vis} = Softmax\left(\frac{Q_{thz}K_{vis}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot V_{vis} \quad (9)$$

The final fused representation is obtained by concatenating these cross-attention features and projecting them back to the original dimension via a Multi-Layer Perceptron (MLP), yielding a unified teacher embedding for distillation [35]:

$$f_{fused} = MLP([T_{vis}\to thz \parallel T_{thz}\to vis]) \quad (10)$$

This fused representation effectively combines the semantic robustness of the visible domain with the modality-specific geometry of the THz domain, serving as a comprehensive target for the student encoder [36].

2.4.3. Teacher–Student Knowledge Distillation

To ensure that the final deployed system operates exclusively on THz dataset, we incorporate a teacher–student distillation framework. The teacher branch leverages the fused THz–visible embedding f_{fused}, along with its task-specific heads for identity verification and head-posture prediction (Figure 3). In parallel, the student branch relies solely on the THz-derived embedding f_{thz}, extracted from the THz encoder E_{thz}, and uses analogous prediction heads. During training, knowledge is transferred from teacher to student through two complementary distillation objectives, enabling the THz-only student model to learn semantic structure and discriminative capability from the multi-modal teacher model [25].

At the representation level, the student embedding f_{thz} is projected into a shared latent space and trained to match the teacher's fused embedding [37]. This is formulated as

$$\phi_S(f_{thz}) = Head_{ver}^{Student}(f_{thz}), \phi_T(f_{fused}) = Head_{ver}^{Teacher}(f_{fused}) \quad (11)$$

$$L_{\text{feature\_distillation}} = \frac{1}{N}\sum_{i=1}^{N}\left(\phi_S(f_{thz})_i - \phi_T(f_{fused})_i\right)^2 \quad (12)$$

 doi:10.20944/preprints202601.1893.v1

The cost function computes the squared L2 distance between student and teacher embeddings. f_{thz} is the latent feature map produced by the THz encoder E_{thz}. ϕ_S is a student projection head from the encoder output into an embedding vector [38]. f_{fused} is the multi-modal fused latent feature map created by integrating THz and visible features via cross-attention (Eq. 10), and ϕ_T is the teacher projection head, producing the target representation.

The feature-level distillation objective encourages the THz encoder to absorb semantic structure derived from the visible modality, including identity-discriminative cues, posture-related information, and both global and local relational patterns encoded in the fused representation [23]. Since the teacher integrates multi-modal inputs, its fused embeddings provide a richer and more informative target than THz-only features. By aligning with this target, the student effectively approximates modality-specific information [39] unavailable in the THz domain, thereby enhancing its representational capacity while remaining single modality at inference.

For posture classification, we employ a logit-level distillation objective in which the student is supervised by the teacher's softened output distribution [40]. The loss is defined as

$$L_{\text{logit\_distillation}} = \tau^2 \cdot KL(\phi_T \parallel \phi_S) \tag{13}$$

$$KL(\phi_T \parallel \phi_S) = \sum_i softmax\left(p_{T,i}/\tau\right) log\left(\frac{softmax\left(p_{T,i}/\tau\right)}{softmax\left(p_{S,i}/\tau\right)}\right) \tag{14}$$

where p_T and p_S denote the teacher and student logits, respectively, and τ is a temperature parameter that smooths the probability distributions. p_T are the raw, pre-Softmax outputs of the teacher classifier, that encode the teacher's confidence for each posture class before normalization. p_S are the corresponding raw outputs of the student (THz-only) classifier. τ is the temperature scaling parameter which smooths the probability vector distribution for a value greater than 1.0 [40]. Both teacher and student logits are passed through a Softmax, with the temperature scaling parameter. The KL divergence (Eq. 14) measures how different the student's softened distribution is from the teacher's. This formulation provides richer inter-class similarity information than hard labels, allowing the THz-only classifier to approximate the more informative and discriminative decision boundaries learned by the fused multi-modal teacher. The τ^2 scaling factor is a standard correction factor in temperature-based distillation. It preserves gradient magnitudes when increasing the temperature, otherwise raising τ would artificially shrink gradients, weakening training. By matching the teacher's softened logits, the student learns fine-grained inter-class relationships and smoother decision boundaries, while implicitly absorbing multi-modal cues present in the fused teacher representation. This enables the THz-only classifier to benefit from the behavior of the more expressive multi-modal teacher during inference.

During training, the model processes fully paired multi-modal inputs: a THz triplet (concealed anchor, unconcealed positive, unconcealed negative) is fed into the THz encoder, while a corresponding visible triplet (concealed anchor, unconcealed positive, unconcealed negative) is processed by the visible encoder. These parallel feature streams are integrated via a symmetric dual cross-attention mechanism, where the visible triplet queries the THz features for non-visual geometric context, and the THz triplet queries the visible features to hallucinate missing textural details. Crucially, the visible stream serves strictly as privileged information [41]. Once training is complete, the entire teacher pathway that comprises the visible encoder, the cross-attention fusion module, and all teacher-specific prediction heads are discarded. At inference, only the THz encoder and the student's verification and head-posture classification are retained. This ensures that the deployed system maintains the computational efficiency of a single-modality THz model while leveraging the robust, structure-aware representations distilled from the multi-modal teacher. The Multi-Modal Teacher-Student training framework versus the THz-only model at inference are described in Figure 5.

 doi:10.20944/preprints202601.1893.v1

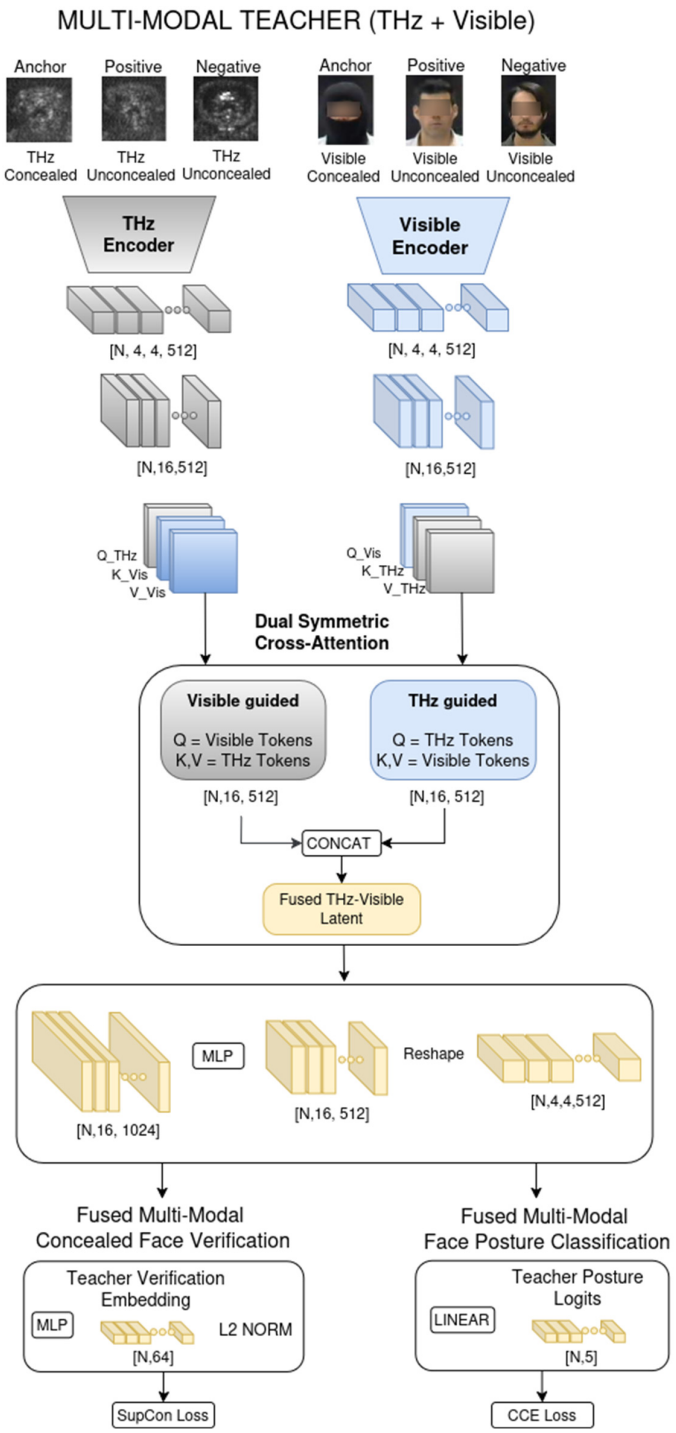


Figure 4. Multi-Modal Teacher Architecture. The framework employs a Dual-Encoder design where Visible and THz encoders process input modalities in parallel. A Dual Cross-Attention module symmetrically fuses features, allowing each modality to query complementary information from the other. The resulting tokens are concatenated and projected into a unified Teacher Embedding, which drives task-specific heads for identity verification and facial posture classification. This fused representation serves as the semantic target for student distillation.

 doi:10.20944/preprints202601.1893.v1

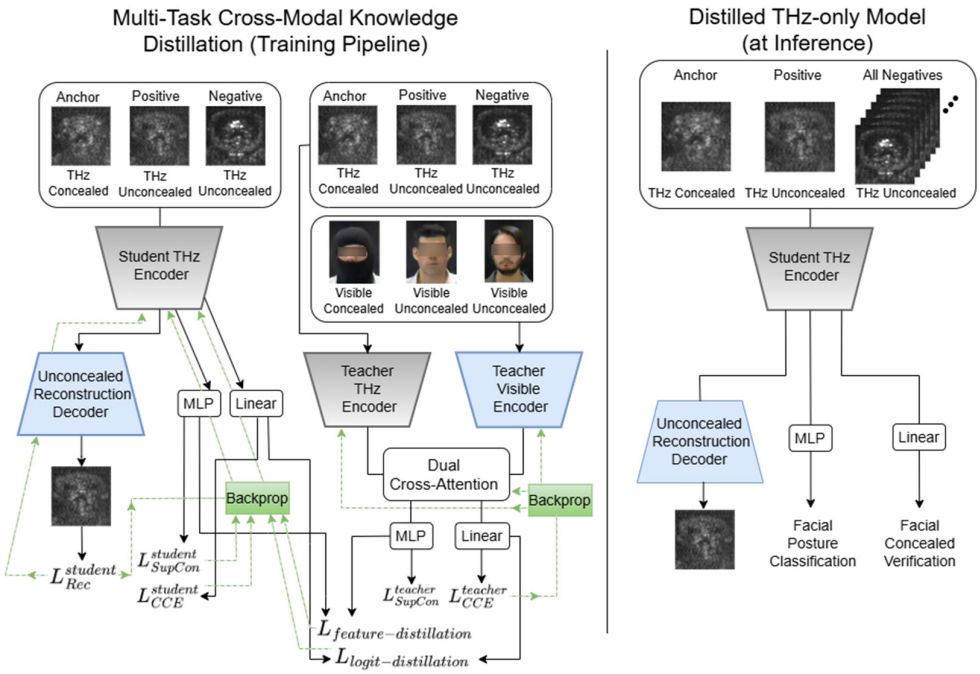


Figure 5. Training and inference pipelines in the Cross-Modal Teacher–Student Knowledge Distillation framework. During training (left part), student and teacher models are optimized jointly through knowledge distillation and their own loss terms. The green dashed lines indicate the back-propagation direction for each loss term. Distillation losses are computed between the verification embeddings and classification logits of the teacher and student networks, and are used to optimize the student model parameters. At inference (right part), the teacher is discarded and predictions are generated using only the student model with THz input data.

3. Results

The dataset was prepared from a total of approximately 1400 images. We employed only samples where for a given subject and facial posture, both concealed and unconcealed images were available. This was necessary for the triplet-based training methodology, which requires anchor (concealed) and positive (unconcealed) pairs. This selection process resulted in a usable dataset of 1158 images.

This dataset was then divided into training/validation and testing sets using a sterile, stratified methodology. This approach guarantees that there is no overlap of a subject's specific image samples between the sets, preventing any form of data leakage that could inflate performance metrics. The stratification was performed based on both the subject's identity and the specific facial posture, ensuring that all variations were proportionally represented.

3.1. Concealed THz Face Identity Verification

The quantitative analysis of the concealed face verification task reveals that both the THz baseline (Model A) and the distilled multi-modal model (Model B) achieve effective performance through distinct learning strategies. A closer inquiry into the results uncovers a trade-off between latent space compactness and decision boundary separation.

The baseline Model A, utilizing only THz images, achieves a slightly higher accuracy (96.08%) compared to the distilled Model B (94.12%). This suggests that Model A successfully optimizes discriminative features within the constraints of the single modality. However, the latent space structure highlights the different methodologies employed by each architecture. Model A prioritizes maximal intra-class compactness, resulting in an average positive-pair distance near zero (0.0540). Model B exhibits a slightly larger average positive distance (0.1028). While less compact than the

 doi:10.20944/preprints202601.1893.v1

baseline, this value remains low (~0.1), suggesting that the distilled embeddings may be preserving natural semantic variations such as facial posture, transferred from the visible-spectrum teacher during the distillation process. Regarding inter-class separation, the distilled Model B demonstrates an advantage, with an average negative distance of 0.7891, which is approximately 6% higher than the baseline's 0.7459. By creating a wider margin between different identities, Model B establishes a broader separation for verification decisions, which makes it slightly more robust as a biometric system [42,43].

Table 1. Concealed face verification results. Quantitative comparison between Model A (THz-only baseline) and Model B (distilled model).

Metric	Model A (THz Baseline)	Model B (Distilled Multi-Modal)	Insight & Explanation
Accuracy	96.08%	94.12%	
Avg. Positive Distance	0.0540	0.1028	Model A achieves better intra-class compactness, whereas Model B's larger positive distance still shows a high intra-class compactness even though it was distilled with visible facial images through distillation process.
Avg. Negative Distance	0.7459	0.7891	By prioritizing inter-class separation, the distilled model (B) has a higher distance between different identities by over 6%, creating a slightly wider margin for verification decisions compared to the baseline's (A) more compact representation.

3.2. Facial Posture Classification Decoder

In the facial posture classification task (Table 2), both the baseline and distilled architectures achieved 100% accuracy on the test set. This indicates that the feature representations learned by both Model A and Model B are sufficient to perfectly distinguish subject postures. A closer examination of the output probabilities reveals a nuance in decision certainty. Model B exhibits a slightly higher Average Correct Confidence (0.8866) compared to the baseline (0.8604). This marginal increase suggests that the cross-modal distillation process effectively improved probability calibration, guiding the model to assign higher likelihoods to the ground-truth classes without compromising the overall stability established by the baseline.

Table 2. Facial posture classification results. Quantitative comparison between Model A (THz-only baseline) and Model B (distilled model).

Metric	Model A (THz Baseline)	Model B (Distilled Multi-Modal)	Insight & Explanation
Accuracy	100.0%	100.0%	
Average Correct Confidence	0.8604	0.8866	The distilled model assigns a little higher probability to the ground-truth class, reflecting somewhat better calibration through teacher guidance.

3.3. Unconcealed Face Reconstruction Decoder

Table 3 presents the reconstruction quantitative quality metrics of the reconstructed unconcealed faces against the ground truth for both, the THz baseline (Model A) and the distilled multi-modal model (Model B). A visual comparison of the reconstructed unconcealed faces against the ground truth for both models is provided in Figure 5. The quantitative metrics reveal statistically equivalent performances of both models. Model A achieves a Mean SSIM of 0.4833 and Mean PSNR of 24.26 dB, while Model B follows closely with a Mean SSIM of 0.4790 and Mean PSNR of 24.18 dB. The weakest-

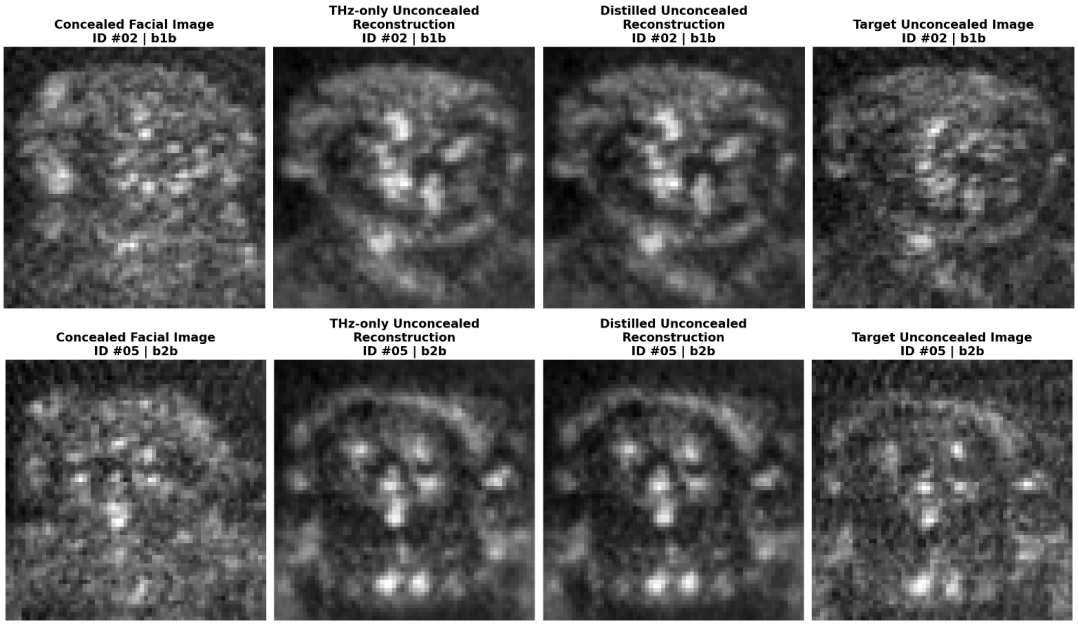
 doi:10.20944/preprints202601.1893.v1

case performance, observed in the frontal view, also remains nearly identical (SSIM ~0.44) for both models.

A qualitative visual inspection of Figure 5 offers further insight beyond these numerical metrics. Visually, the reconstructed unconcealed THz images from both models appear robust, successfully preserving key facial structures and biometric patterns inherent to the subjects. The generated output exhibits a smooth denoised appearance. This smoothing results in coherent facial representations that allow for clearer interpretation of facial morphology, suggesting that the perceptual quality of the reconstruction may exceed what strict pixel-level metrics like SSIM imply. This performance parity is significant as it demonstrates that the distillation process did not compromise the generative capabilities of the network. Typically, optimizing an embedding space for a discriminative task (verification), risks "feature collapse" [44] where the fine-grained textural information needed for reconstruction is discarded in favor of abstract, identity-focused representations. However, the results indicate that the multi-modal distillation effectively regularized the latent space without such degradation. Model B maintains high generative fidelity comparable to the baseline, confirming that the shared weights successfully accommodate the gradients from the reconstruction loss, even while simultaneously adapting to the cross-modal supervision required for the verification and classification tasks.

Table 3. Unconcealed face reconstruction results. Quantitative comparison between Model A (THz-only baseline) and Model B (distilled model). Statistically equivalent performances of both models were obtained for this task.

Metric	Model A (THz Baseline)	Model B (Distilled Multi-Modal)
Mean SSIM (Reconstructed vs. GT)	0.4833	0.4790
Mean PSNR (Reconstructed vs. GT)	24.26 dB	24.18 dB
SSIM Weakest Case (Front View)	0.4441	0.4429



 doi:10.20944/preprints202601.1893.v1

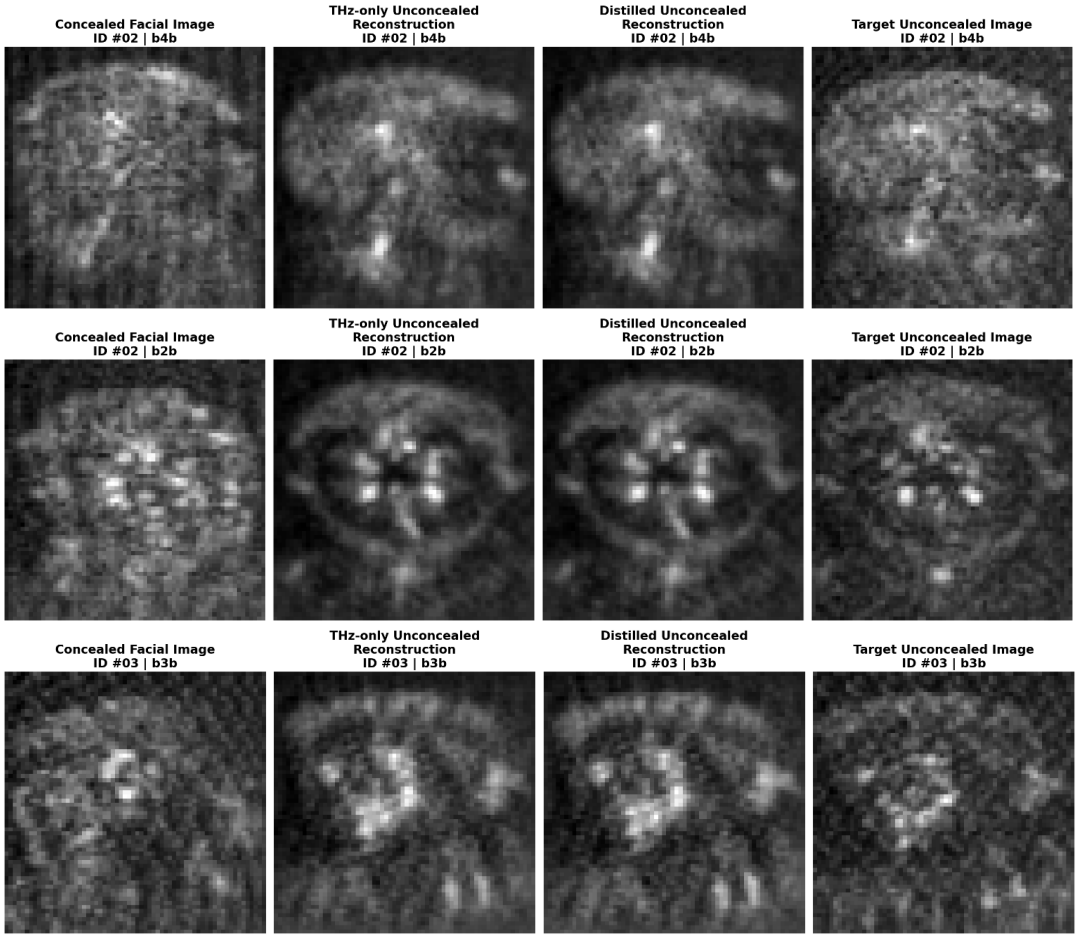


Figure 6. A visual comparison of unconcealed face reconstruction from concealed THz images for different identities and postures. From left to right: (a) concealed facial input images, (b) THz-only model reconstructions, (c) distilled model reconstructions, and (d) ground truth unconcealed target images. Visual inspection reveals comparable reconstruction fidelity between the two models. Subject identifiers and posture codes are indicated above each image.

4. Discussion

This study introduces and evaluates a unified Multi-Task Learning (MTL) framework for concealed face analysis in the THz domain. By comparing a single THz modality baseline (Model A) with a multi-modal (THz and visual) distilled approach (Model B), we uncover insights into the mechanisms of learning from sparse, non-literal sensory data.

The comparative analysis across the three tasks illustrates distinct optimization paths for each architecture. While the single-modality baseline achieves better latent space compactness, the cross-modal distillation approach shows higher inter-class separation and semantic generalization. This suggests that teacher-guided supervision effectively alters the internal representation, enhancing decision margins and probability calibration without compromising the model's capacity for high-fidelity generative reconstruction.

While the THz-only model (Model A) attained a slightly superior verification accuracy of 96.08% and exhibited excellent positive-class compactness (Avg Positive Distance ≈ 0.0540), this performance highlights its specific optimization strategy. The results indicate that the model has successfully structured a highly specialized feature space within the THz domain, prioritizing tight intra-class clustering to maximize discriminative power within the constraints of the single modality.

The distilled Model B has a latent structure characterized by larger positive distances (~0.10) and increased inter-class separation (~0.79). These findings highlight the importance of evaluating

 doi:10.20944/preprints202601.1893.v1

biometric systems not just on raw accuracy, but also on latent space topology, as these margins offer insights into how multi-modal constraints shape system robustness.

Both architectures achieved perfect facial posture classification accuracy. The distilled Model B exhibited a slightly higher average confidence in its correct predictions (0.89) compared to the baseline. This difference implies that the knowledge transfer process potentially refines the calibration of the output probabilities. In the context of THz imaging, where visual cues can be sparse, this enhancement suggests that the student model effectively leverages the semantic guidance from the teacher.

The reconstruction analysis reveals that the distilled model maintained generative fidelity quantitatively indistinguishable from the baseline, despite the added verification constraints. This indicates that the multi-task objective successfully preserved essential morphological details, avoiding feature collapse often associated with discriminative optimization. Consequently, the results support the viability of unified architectures where a single encoder efficiently handles both robust verification and interpretable reconstruction.

These results align with recent literature advocating for the potential of incorporating richer semantic-aware feature learning in sparse modalities [41,45].

5. Conclusions

This work establishes a unified multi-task framework that successfully delivers novel objectives in THz facial imaging, that include concealed face verification, facial posture classification, and generative reconstruction of unconcealed faces from concealed faces on a very challenging THz image dataset. We show that robust multi-objective performance is attainable even when relying exclusively on a THz-only dataset. Beyond performance metrics, this study demonstrates that cross-modal distillation can effectively regularize a student model without the need for paired multi-spectral (visible-THz) data at inference time. This may provide a reference for deploying high-fidelity biometric systems in resource-constrained environments, potentially enhancing decision boundaries and uncertainty calibration. Our findings illustrate the feasibility of unified architectures that do not compromise generative interpretability for the sake of discriminative accuracy. The ability of a single multi-task encoder to support both secure verification and human-readable reconstruction suggests a path toward more efficient, privacy-preserving security scanners. Future research can further explore adaptive distillation mechanisms, potentially leading to hybrid systems that blend the structural precision of optical imaging with the unique penetrative capabilities of sub-millimeter waves.

Author Contributions: Conceptualization, N.B. and Y.Y.; methodology, N.B. ,Y.Y; software, N.B.; validation, N.B.; formal analysis, N.B.; investigation, N.B., I.O.Y., A.B.S., H.A.; resources, N.B., I.O.Y., A.B.S., H.A.; data curation, N.B., I.O.Y., A.B.S., H.A.; writing—original draft preparation, N.B.; writing—review and editing, N.B. ,Y.Y.; visualization, N.B.; supervision, Y.Y.; project administration, Y.Y.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Institutional: Review Board Statement Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The complete implementation code and trained weights are provided at https://github.com/noamberg/thz-face-distil. The facial dataset is currently not publicly available.

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

 doi:10.20944/preprints202601.1893.v1

1. Hao, J.; Li, J.; Pi, Y. Three-dimensional imaging of terahertz circular SAR with sparse linear array. Sensors 2018, 18, 2477.
2. Amini, T.; Jahangiri, F.; Ameri, Z.; Hemmatian, M.A. A review of feasible applications of THz waves in medical diagnostics and treatments. J. Lasers Med. Sci. 2021, 12, e92.
3. Tao, Y.H.; Fitzgerald, A.J.; Wallace, V.P. Non-contact, non-destructive testing in various industrial sectors with terahertz technology. Sensors 2020, 20, 712.
4. Phing, S.H.; Mazhorova, A.; Shalabi, M.; Peccianti, M.; Clerici, M.; Pasquazi, A.; Ozturk, J.A. Sub-wavelength terahertz beam profiling of a THz source via an all-optical knife-edge technique. Sci. Rep. 2015, 5, 8551.
5. Yildirim, I.O.; Altan, H.; Şahin, A.B. Performance of an active THz imaging system for recognition of concealed faces. J. Infrared Millim. Terahertz Waves 2023, 44, 365-378.
6. Hashemi, N.S.; Aghdam, R.B.; Ghiasi, A.S.B.; Fatemi, P. Template matching advances and applications in image analysis. arXiv 2016, arXiv:1610.07231.
7. Sharma, S. Template matching approach for face recognition system. Int. J. Signal Process. Syst. 2013, 1, 284-289.
8. Brunelli, R.; Poggio, T. Face recognition: Features versus templates. IEEE Trans. Pattern Anal. Mach. Intell. 1993, 15, 1042-1052.
9. Hassner, T.; Harel, S.; Paz, E.; Enbar, R. Effective face frontalization in unconstrained images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7-12 June 2015; pp. 4295-4304.
10. Gonzalez-Sosa, E.; Vera-Rodriguez, R.; Fierrez, J.; Patel, V.M. Millimetre wave person recognition: Hand-crafted vs learned features. In Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), New Delhi, India, 22 February 2017.
11. Su, W.T.; Hung, Y.C.; Yu, P.J.; Yang, S.H.; Lin, C.W. Seeing through a black box: toward high-quality terahertz imaging via subspace-and-attention guided restoration. In Proceedings of the European Conference on Computer Vision, Tel-Aviv, Israel, 23-27 October 2022.
12. Alefs, B.G.; Den Hollander, R.J.M.; Nennie, F.A.; Van Der Houwen, E.H.; Bruijn, M.; Van Der Mark, W.; Noordam, J.C. Thorax biometrics from millimetre-wave images. Pattern Recognit. Lett. 2010, 31, 2357-2363.
13. Moreno-Moreno, M.; Fierrez, J.; Vera-Rodriguez, R.; Parron, J. Distance-based feature extraction for biometric recognition of millimeter wave body images. In Proceedings of the Carnahan Conference on Security Technology, Barcelona, Spain, 18-21 October 2011.
14. Palka, N.; Kowalski, M. Towards fingerprint spoofing detection in the terahertz range. Sensors 2020, 20, 3379.
15. Gonzalez-Sosa, E.; Vera-Rodriguez, R.; Fierrez, J.; Alonso-Fernandez, F.; Patel, V.M. Exploring body texture from mmW images for person recognition. IEEE Trans. Biom. Behav. Identity Sci. 2019, 1, 139-151.
16. Zeng, Z.; Wu, H.; Chen, M.; Luo, S.; He, C. Concealed hazardous object detection for terahertz images with cross-feature fusion transformer. Opt. Lasers Eng. 2024, 182, 108454
17. Jayaweera, S.S.; Regani, S.D.; Hu, Y.; Wang, B.; Ray Liu, K.J. mmID: High-resolution mmWave imaging for human identification. In Proceedings of the 2023 IEEE 9th World Forum on Internet of Things (WF-IoT), Aveiro, Portugal, 12-27 October 2023; pp. 1-6.
18. Ge, Z.; et al. Deep-learning-based method for concealed object detection in terahertz (THz) images. In Advanced Fiber Laser Conference (AFL2023); SPIE: Bellingham, WA, USA, 2024; pp. 268-274.
19. Katsuyama, Y.; Sato, T.; Qi, X.; Tamesue, K.; Wen, Z.; Yu, K.; Tokuda, K.; Sato, T. Deep learning based concealed object recognition in active millimeter wave imaging. In Proceedings of the 2021 IEEE Asia-Pacific Microwave Conference (APMC), Brisbane, Australia, 28 November-1 December 2021; pp. 434-436.
20. Cheng, L.; Ji, Y.; Li, C.; Liu, X.; Fang, G. Improved SSD network for fast concealed object detection and recognition in passive terahertz security images. Sci. Rep. 2022, 12, 12082.
21. Xing, W.; Zhang, J.; Guo, L. A fast detection method based on deep learning of millimeter wave human image. In Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality, Takamatsu, Japan, 23-25 November 2018; pp. 67-71.

 doi:10.20944/preprints202601.1893.v1

22. Peng, D.; Xu, L.; Wu, H.; Wang, T.; Xiao, H.; Cheng, L.; Qin, Y. Multi-scale super-resolution reconstruction of terahertz images for postal security inspection. Opt. Express 2025, 33, 16237-16252.
23. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-modality person re-identification with shared-specific feature transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020; pp. 13379-13389.
24. Ranjan, R.; Patel, V.M.; Chellappa, R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 41, 121-135.
25. Shan, L.; Zhang, R.; Chilukoti, S.V.; Zhang, X.; Lee, I.; Hei, X. IdentityKD: Identity-wise Cross-modal Knowledge Distillation for Person Recognition via mmWave Radar Sensors. In Proceedings of the 6th ACM International Conference on Multimedia in Asia, Bangkok, Thailand, 3-6 December 2024; pp. 1-7.
26. Yip, B.; Towner, R.; Kling, T.; Chen, C.; Wang, Y. Image pre-processing using OpenCV library on MORPH-II face database. arXiv 2018, arXiv:1811.06934.
27. Wang, F.; Xiang, X.; Cheng, J.; Yuille, A.L. Normface: L2 hypersphere embedding for face verification. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23-27 October 2017; pp. 1041-1049.
28. Khosla, P.; et al. Supervised contrastive learning. Adv. Neural Inf. Process. Syst. 2020, 33, 18661-18673.
29. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 2472-2481.
30. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 586-595.
31. Li, H.; Wu, X.J. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. Inf. Fusion 2024, 103, 102147.
32. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. Smil: Multimodal learning with severely missing modality. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2-9 February 2021; Volume 35, pp. 2302-2310.
33. Yu, Z.; Wang, J.; Yu, L.C.; Zhang, X. Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual, November 2022; pp. 414-423.
34. Thoker, F.M.; Gall, J. Cross-modal knowledge distillation for action recognition. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22-25 September 2019.
35. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. Adv. Neural Inf. Process. Syst. 2021, 34, 14200–14213.
36. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 345–360.
37. Garcia, N.C.; Morerio, P.; Murino, V. Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8-14 September 2018.
38. Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; Choi, J.Y. A comprehensive overhaul of feature distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1921–1930.
39. Kowalski, M.; Grudzień, A.; Mierzejewski, K. Thermal–visible face recognition based on cnn features and triple triplet configuration for on-the-move identity verification. Sensors 2022, 22, 5012.
40. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. arXiv 2015, arXiv:1503.02531.
41. Tong, Q.; Nocentini, O.; Lagomarsino, M.; Cai, K.; Lorenzini, M.; Ajoudani, A. Lightweight Facial Landmark Detection in Thermal Images via Multi-Level Cross-Modal Knowledge Transfer. arXiv 2025, arXiv:2510.11128.

 doi:10.20944/preprints202601.1893.v1

42. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7-12 June 2015; pp. 815-823.
43. Lu, J.; Hu, J.; Tan, Y.P. Discriminative deep metric learning for face and kinship verification. IEEE Trans. Image Process. 2017, 26, 4269-4282.
44. Xue, Y.; Joshi, S.; Gan, E.; Chen, P.Y.; Mirzasoleiman, B. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23-29 July 2023; pp. 38938-38970.
45. Hafner, F.M.; Bhuyian, A.; Kooij, J.F.; Granger, E. Cross-modal distillation for RGB-depth person re-identification. Comput. Vis. Image Underst. 2022, 216, 103352.