Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Unifying Thermal and Visible Face Recognition Through Continuous Convolution Global Weighting Transformers and Fourier Neural Operators

Areej A. Abed[1a*], Abdul Monem S.Rahma[2b], Omar A. Dawood[3c]

[1] Computer Science Department College of Computer Science and Information Technology, University of Anbar, AL-Anbar, Iraq,*

[2] Computer Science Department, Al-Mansour University college, Baghdad, Iraq

[3]Computer Science Department, College of Computer Science and Information Technology, University of Anbar, AL-Anbar, Iraq

Authors Emails

a) are21c1004@uoanbar.edu.iq.

b) Monem.rahma@muc.edu.iq

c) omar-abdulrahman@uoanbar.edu.iq

ARTICLE INFO

Received: 25 Dec 2024

Revised: 15 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Cross-spectrum face recognition, in which thermal and visible images must be jointly analyzed, has long been challenged by discrepancies in illumination, sensor noise, and spectral characteristics. These issues are particularly relevant in security, defense, and healthcare, where robust identification across lighting and environmental conditions is essential. Despite advances in standard convolutional or attention-based networks, many models still struggle with domain adaptation and fail to extract consistent features from thermal and visible inputs. To address this gap, we investigate three alternative architectures: a Continuous Neural Network (CNN) that learns smooth kernel functions, an Attention-Free Transformer (AFT) with global weighting instead of multi-head attention, and a Fourier Neural Operator (FNO) that operates on low-frequency spectral components. Each model was trained on a disjoint set of thermal–visible face images and then evaluated for classification accuracy. Whereas the FNO-based method reached 0.86 (0.85 macro-average F1-scores), our results reveal that the continuous neural network and the attention-free transformer attained 0.98 accuracies (with macro- and weighted-average F1-scores of 0.98).

Keywords: cross-spectrum; thermal–visible fusion; face recognition; Continuous Convolution Networks; Attention-Free Transformers; Fourier Neural Operator

Introduction

Face recognition and identification in different light conditions are still difficult problems, especially when matching different spectra of face images [1]. Particularly, thermal-visible face identification requires bridging significant modality gaps: Thermal images convey robust performance in poor lighting but typically lack the discriminative appearance features available in visible data. On the other hand, visible images convey detailed color and texture information, which deteriorates significantly in poor light conditions [2]. Building a robust cross-spectrum face identification system is crucial in many fields, including security, surveillance cameras, and healthcare, where identification mostly occurs in uncontrolled or poor lighting conditions [3].

In the last few years, many earlier approaches to processing the spectral mismatch achieved high performance depending on Deep Learning (DL) methods, specifically Convolutional Neural Networks (CNNs) or transformer-based architectures. While CNN-based pipelines have shown promise in inhomogeneous fields, they fail to align thermal and visible modalities without large-scale adaptation or domain-specific engineering [4]. On the other hand, transformer models can model global relations, but multi-head self-attention is computationally expensive and prone to overfitting, particularly with smaller cross-spectrum datasets. Thus, new strategies are needed to mitigate computational overhead and perform flexible, domain-adaptive feature extraction [5].

This paper explores three alternative deep architectures to alleviate these challenges, each offering a different strategy for bridging thermal–visible domain gaps. In particular, continuous neural networks (CNN-based models parameterizing their convolutional filters as smooth, learnable functions) have shown promise on tasks involving extreme local appearance shifts as they modulate their kernels to more subtly adapt to domain-specific patterns [6]. Alternatively, attention mechanisms—originally popularized by the vision transformer—have proven effective at global context modeling for unimodal tasks [7]. However, the overhead of multi-head self-attention may be unwanted, and some evidence indicates it is not necessarily needed to model cross-domain cues [8]. An attention-free transformer (AFT) is a newer alternative that forgoes multi-head projections for a simpler global weighting, simplifying model complexity without compromising salient global interactions [9].

The Fourier neural operator (FNO) has also gained traction as a spectral approach, where low-frequency components undergo learnable transformations in the frequency domain before being inverse-transformed back to spatial space [10]. This process can integrate cross-spectrum face features by aligning thermal and visible signals at the spectral level, an idea supported by promising results in image-to-image translation tasks [11]. However, FNO-based methods sometimes underperform when the sensor gap is especially large, as high-frequency details unique to each domain may be truncated [12]. Consequently, each approach—continuous convolutions, attention-free transformers, and Fourier operators—exhibits both strengths and limitations when facing thermal–visible matching scenarios.

In this work, we systematically evaluate and compare three architectures for cross-spectrum face recognition:

1. Continuous Neural Networks (CNNs), which adapt convolutional filters via smooth parameterization,

2. Attention-Free Transformers (AFTs), leveraging a global weighting mechanism that discards multi-head complexity, and

3. Fourier Neural Operator (FNO) blocks focusing on low-frequency spectral alignment.

Every model is evaluated on a cross-spectrum face dataset with twenty identity classes. With strictly disjoint train/validation splits. Our contributions thus lie in proposing specialized cross-spectrum architectures, delivering rigorous comparisons, and demonstrating the viability of alternative neural operators beyond classical CNNs or standard transformers for robust thermal–visible face recognition.

This research makes the following contributions:

1. Propose and unify three architectures (continuous convolution, attention-free transformer, and FNO) for cross-spectrum face recognition, providing a multifaceted view of how best to integrate disparate sensor data.

2. Present comprehensive comparative results, where the continuous neural network and attention-free transformer attain top-tier accuracy (98%) and the FNO-based method

achieves 86% accuracy. These findings reveal the trade-offs in complexity, computational load, and sensitivity to high-frequency information.

3. Offer practical insights for domain-specific improvements: continuous kernel parameterization can yield adaptable filters, global weighting can simplify attention overhead, and spectral-domain alignment can robustly unify low-frequency content despite modality variations.

In the following sections, we detail the theoretical underpinnings of each proposed network, describe our thermal–visible dataset and training protocols, and analyze the numerical outcomes in terms of accuracy, F1-scores, confusion matrices, and multi-class Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. Our results underscore that, although each approach addresses cross-spectrum disparities in its own manner, both the continuous and attention-free solutions demonstrate near-ideal performance. In contrast, the Fourier-based approach, while promising, faces greater challenges in preserving high-frequency cues across modalities. By illuminating these strengths and weaknesses, we provide researchers and practitioners with targeted guidance on deploying robust cross-spectrum solutions in real-world scenarios.

Related Work

Cross-spectrum face recognition, particularly in thermal-to-visible matching, has seen significant advances due to deep learning [12]. However, domain adaptation, spectral discrepancies, and feature preservation challenges persist. Below, we review recent efforts in this area.

Deep learning-based methods to close the spectral gap have been investigated in several papers. In [13] a Deep Joint Independent Component Analysis Network (DJICAN) was developed to improve feature alignment by learning mappings between thermal and visual face images. Similarly, [2] introduced a Domain and Pose Invariant Framework to address pose variations and spectral discrepancies. Their work improved matching accuracy under extreme conditions.

Recent efforts have also focused on generative approaches. [13] proposed a Denoising Diffusion Probabilistic Model (DDPM) for Thermal-to-Visible (T2V) image translation, achieving state-of-the-art results. Similarly, [13] utilized Conditional GANs (CGANs) to generate thermal face images from visible ones, enhancing recognition performance.

Transformer-based methods have also been explored. [3] investigated end-to-end deep learning solutions for thermal spectrum face verification, addressing the limitations of prior CNN-based methods. Additionally, [13] reviewed the latest deep infrared (IR) approaches, identifying key challenges in spectral fusion. Table 1 summarizes key recent works, their methodologies, and the gaps our research aims to address.

However, while these efforts show promise, they do not systematically compare different architectural paradigms tailored to cross-spectrum face recognition. Our work fills this gap by evaluating three distinct architectures:

1. Continuous convolution networks, which adapt kernel filters for spectral alignment.

2. Attention-free transformers, reducing complexity while retaining global feature interactions.

3. Fourier Neural Operators (FNOs), leveraging spectral domain transformations to align thermal-visible face signals.

Table 1: Recent works, their methodologies, and the gaps our research aims to address.

Study	Methodology	Research Gap
[13] (2022)	Independent Component Analysis for spectral alignment.	Lacks adaptation to large spectral gaps.
[2] (2022)	Domain and pose-invariant framework.	Does not address computational overhead in transformers.
[14] (2022)	Diffusion-based image synthesis for T2V translation.	Computationally expensive and slow inference.
[15] (2022)	Generative adversarial networks for thermal-to-visible translation.	Struggles with fine-grained identity retention.
[16] (2022)	YOLOv5-based face and landmark detection in thermal images.	Not designed for full face recognition tasks.
[17] (2022)	Cross-modality discriminator network.	Lacks frequency-based feature extraction.
[4] (2022)	Triple CNN architecture for thermal-visible face verification.	Limited performance on extreme spectral shifts.
[18] (2022)	GAN-based thermal image generation for robust face recognition.	May introduce artifacts affecting recognition accuracy.
[19] (2023)	Survey on deep learning-based IR face recognition.	Does not compare transformer vs. Fourier-based methods.
[3] (2024)	End-to-end CNN for face verification.	Lacks attention-free and spectral-domain methods.

Our study compares three alternative neural network architectures: continuous convolution networks, attention-free transformers, and Fourier neural operators. Unlike previous research, which often focuses on a single approach, we systematically evaluate distinct architectural paradigms for cross-spectrum face recognition. Specifically, to reduce the model complexity while maintaining robust performance, we explored the AFT architecture in global feature weighting. Furthermore, a novel image alignment on a thermal-visible image perspective is offered using FNOs. Finally, multiple performance metrics were used on a controlled dataset to ensure a comprehensive assessment of the results, including accuracy, F1-score, recall, sensitivity, ROC curves, and confusion matrix.

Methods

This study addresses the challenges in thermal-visible face identification by proposing an end-to-end three distinct architectures that consolidate three distinct architectures, each emphasizing a specific strategy for bridging the modality gap. The continuous convolution network leverages adaptive kernel parameterization to accommodate illumination and spectral detail disparities. The Attention-Free Transformer (AFT) dispenses with multi-head self-attention in favor of a global weighting mechanism, thereby capturing broad contextual cues without overwhelming computational resources. By contrast, the Fourier Neural Operator (FNO) operates in the frequency domain to align low-frequency components across thermal and visible images, offering a theoretically elegant way of mitigating spectral mismatches. Each model processes standardized inputs, culminating in a single classification head that outputs identity probabilities. The comparative aim is to determine whether continuously adjustable kernels, globally weighted patch interactions, or spectral alignment best resolve the discrepancy between thermal and visible facial representations in real-world conditions. Figure 1 illustrates the overall pipeline of the proposed method, showing dual-input thermal–visible data preprocessing followed by parallel architectures (Continuous Convolution Network, Attention-Free Transformer, and Fourier Neural Operator) converging into a unified classification head.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

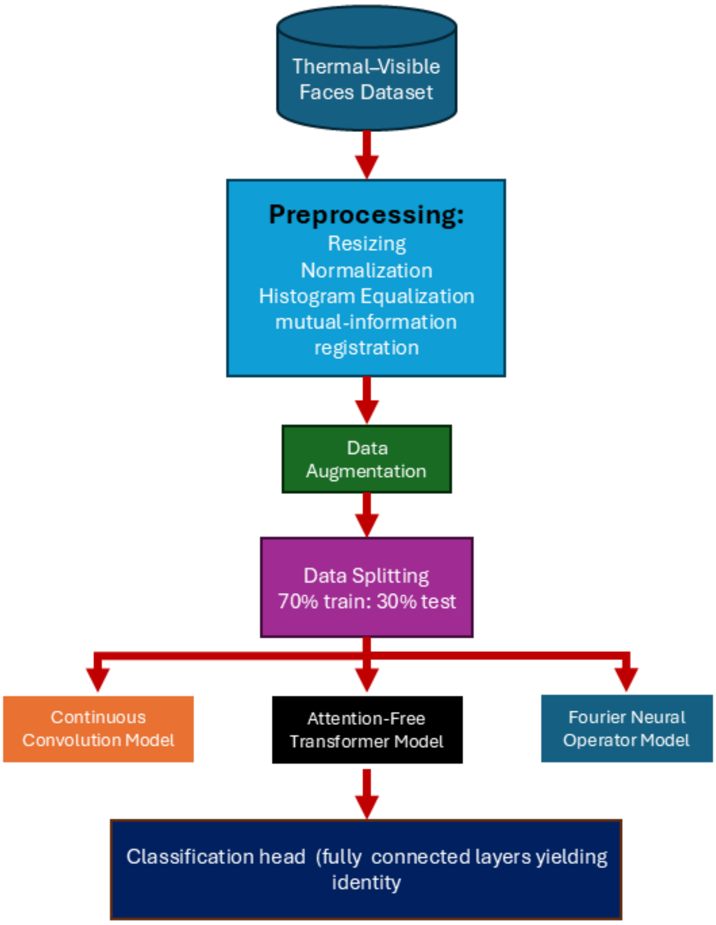

Thermal–Visible Faces Dataset
Preprocessing:
Resizing
Normalization
Histogram Equalization
mutual-information registration
Data Augmentation
Data Splitting
70% train: 30% test
Continuous Convolution Model
Attention-Free Transformer Model
Fourier Neural Operator Model
Classification head (fully connected layers yielding identity

Figure 1: The proposed pipeline for cross-spectrum face recognition.

Data Preprocessing

A structured preprocessing pipeline was used to guarantee correct alignment and effective feature extraction from thermal and visible facial images. Training and validation sets were created from the dataset to guarantee a strong separation between identity classes, thereby preventing data leakage. Two main modalities made up the dataset: thermal and visual images, with different spectral properties.

We applied histogram equalization to improve contrast in thermal images, making features clearer. All images were resized to 224×224 pixels for consistency. We used random cropping, horizontal flipping, Gaussian noise, and Fourier spectrum masking for data augmentation to boost model robustness and generalization. This operation compelled the models to identify and utilize the invariant features under varying frequency distributions, making the models more robust to the spectrum shifts. The pixel intensities were normalized using min-max scaling to maintain data representation consistency, restricting the values uniformly to the range [0,1]. Finally, we registered the thermal and visible image pairs with mutual information-based registration, providing pixel-wise consistency across the modalities.

Continuous Convolution Networks

Traditional convolutional neural networks typically utilize fixed-weight filters, which may prove inadequate in handling the significant spectral shifts that characterize thermal and visible face images. To overcome this limitation, CCNs introduce adaptively parameterized convolutional kernels capable of smoothly and dynamically adjusting to variations in input data. Unlike conventional static-filter approaches that apply uniform kernels across all input images, the CCN architecture facilitates kernel flexibility, thereby enhancing the extraction of common features from diverse image domains. This adaptability is particularly advantageous for cross-spectrum tasks, given that thermal images inherently lack certain high-frequency details commonly present in visible-spectrum images, and conversely, visible images may omit spectral information prominently in thermal imagery.

The final implementation of our continuous convolution network comprises an initial input layer for 224×224×3 images, followed by two continuous convolutional layers, each of which was followed by a rectified linear unit (Relu) activation and a max-pooling operation to shrink spatial dimensions gradually. The tensor is then flattened and sent to a fully connected "penultimate" layer of size 128 neurons before a last dense output layer projects into 20 identity classes. Table 2 offers a layer-by-layer synopsis:

Table 2: Model Summary for the Continuous Convolution Network (CCN)

Layer (type)	**Output Shape**	**Param #**
input_layer (InputLayer)	(None, 224, 224, 3)	0
continuous_conv2d (ContinuousConv2D)	(None, 224, 224, 16)	448
re_lu (ReLU)	(None, 224, 224, 16)	0
max_pooling2d (MaxPooling2D)	(None, 112, 112, 16)	0
continuous_conv2d_1 (ContinuousConv2D)	(None, 112, 112, 32)	4,640
re_lu (ReLU)	(None, 224, 224, 32)	0
max_pooling2d (MaxPooling2D)	(None, 56, 56, 32)	0
flatten (Flatten)	(None, 100352)	0
penultimate_dense (Dense)	(None, 128)	12,845,184
dense (Dense)	(None, 20)	2,580

Total params: 12,852,852

Trainable params: 12,852,852

Non-trainable params: 0

The continuous convolutional filters at the heart of this network are parameterized to accommodate domain-specific distortions, including low-illumination noise in thermal images or overexposed regions in visible images. By learning these smooth, continuous kernels, the model preserves essential identity-relevant information across both spectra more effectively than a standard CNN. At the penultimate stage, a fully connected layer of 128 neurons provides a compact embedding of the extracted features, which are then passed to the final dense layer that generates class probabilities. This multi-stage process supports robust cross-spectrum generalization, allowing the model to distinguish identities even under substantial illumination or spectral disparities.

Attention-Free Transformers

While standard Transformers have demonstrated strong performances on a wide range of vision tasks, their application of multi-head self-attention mechanisms is computationally expensive and prone to overfitting, especially with limited datasets. In response, we created an AFT model that does not employ the standard self-attention but a lighter-weight global weighting mechanism to prevent compromising the long-range feature modeling property of Transformer-based methods but reduces the parameter number and computational complexity by a significant margin.

To lower the input's spatial resolution, our first layers use two-dimensional convolution and max-pooling. The images are subsequently split up and linearly projected to the embedding space. Unlike passing the embeddings through multiple attention heads, the AFT blocks compute global weights and broadcast them to all the patch embeddings, effectively capturing context without the accompanying high computational cost of pairwise attention operations. The model has four stacked AFT blocks, each progressively refining the extracted representations. The blocks conclude with a global average pooling layer, effectively aggregating learned features across patches in the image—finally, a dense layer outputs class probabilities for the target categories of interest.

Table 3 provides a detailed summary of the configuration of each layer, including input and output dimensions, parameter counts, and the functional roles of individual components. By excluding multi-head self-attention, the model effectively reduces the quadratic computational complexity found in traditional Transformer architectures. This design choice makes the AFT model suitable for use in environments with limited resources or applications requiring fast inference capabilities.

Table 3: Model Summary for the Attention-Free Transformer (AFT)

Layer (type)	**Output Shape**	**Param #**
input_layer (InputLayer)	(None, 224, 224, 3)	0
conv2d_6 (Conv2D)	(None, 224, 224, 64)	1,792
max_pooling2d_6 (MaxPooling2D)	(None, 112, 112, 64)	0
conv2d_7 (Conv2D)	(None, 112, 112, 128)	73,856
max_pooling2d_7 (MaxPooling2D)	(None, 56, 56, 128)	0
patch_embedding_3 (PatchEmbedding)	(None, None, 256)	2,097,408
aft_block_14 (AFTBlock)	(None, None, 256)	723,968
aft_block_15 (AFTBlock)	(None, None, 256)	723,968
aft_block_16 (AFTBlock)	(None, None, 256)	723,968
aft_block_17 (AFTBlock)	(None, None, 256)	723,968
global_average_pooling1d_3 (GlobalAveragePooling1D)	(None, 256)	0
dense (Dense)	(None, 20)	5,140

Total params: 5,074,068 (19.36 MB)
Trainable params: 5,074,068 (19.36 MB)
Non-trainable params: 0 (0.00 B)

This architecture prioritizes essential features rather than calculating full attention scores for every image patch, capturing high-level dependencies efficiently. The AFT identifies individuals across thermal and visible spectra without overfitting, providing a computationally efficient solution for cross-spectrum face recognition tasks.

Fourier Neural Operators

Fourier Neural Operators (FNOs) are a new type of deep learning models that operate directly in the frequency domain. Unlike the spatial representation of the features learned by standard CNNs, FNOs perform the operations in the spectrum domain and, therefore, are naturally suitable for cross-spectrum applications where frequency alignment can close modality gaps. Since the thermal and visible spectrum distributions inherently possess discrepancies, we examined the feasibility of FNO as a solution for cross-spectrum face recognition.

The model first applied an FFT to convert input images from the spatial domain to the frequency domain. In this transformed space, a low-pass filtering operation was performed to retain only the dominant frequency components, removing high-frequency noise that could introduce domain discrepancies. The core of the FNO framework consisted of two spectral convolution layers, which learned transformations directly on frequency coefficients. This process allowed the model to align thermal and visible images based on shared frequency components rather than spatial textures. Table 4 below details the layer configuration, parameter counts, and output dimensions.

Table 4: Model Summary for the Fourier Neural Operators (FNOs)

Layer (type)	**Output Shape**	**Param #**
input_layer (InputLayer)	(None, 224, 224, 3)	0
conv2d_36 (Conv2D)	(None, 224, 224, 32)	896
max_pooling2d_6 (MaxPooling2D)	(None, 112, 112, 32)	0
conv2d_7 (Conv2D)	(None, 112, 112, 64)	18,496
max_pooling2d_7 (MaxPooling2D)	(None, 56, 56, 64)	0
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
fno_block_12 (FNOBlock)	(None, 56, 56, 64)	1,086,080
global_average_pooling1d_3 (GlobalAveragePooling1D)	(None, 64)	0
dense (Dense)	(None, 20)	1,300

Total params: 8,709,332 (33.22 MB)
Trainable params: 8,709,332 (33.22 MB)
Non-trainable params: 0 (0.00 B)

Following Fourier domain processing, the network inverted the Fast Fourier Transform (iFFT) to translate spectral data into spatial representations. Identity predictions were generated using a three-layer, completely linked layer of 1024, 512, and 128 neurons, headed by classification. Domain-invariant feature learning was motivated by combining cosine similarity loss with Kullback-Leibler (KL), divergence loss. Over 40 epochs, a batch size of 32 was used in the RMSprop optimizer to train the model.

Performance Metrics

Multiple metrics can provide a comprehensive picture of the quality and resilience of a facial recognition system: accuracy, precision, recall, and the F1-score. All these criteria evaluate model performance differently, particularly in classification tasks.

Accuracy

Accuracy is a fundamental metric that assesses the ratio of properly predicted instances to the total instances, providing a basic indication of the model's performance; nevertheless, it may be misleading in situations including class imbalance. The accuracy is calculated as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \qquad (1)$$

Where:

TP : True Positives (correctly predicted positive instances)

TN: True Negatives (correctly predicted negative instances)

FP: False Positives (incorrectly predicted positive instances)

FN: False Negatives (incorrectly predicted negative instances)

Precision

Precision, also known as Positive Predictive Value, refers to the ratio of true positive predictions to the total number of expected positives. Mathematically, it is a ratio that indicates the number of accurately predicted positive cases relative to all instances projected as positive, rendering accuracy a valuable indicator when the cost of false positives is significant. The precision is calculated as follows:

$$Precision = \frac{TP}{(TP + FP)} \qquad (2)$$

Recall

Recall, also known as Sensitivity or True Positive Rate, is the ratio of anticipated true positive instances to all actual positive cases, indicating the effectiveness of detecting positive instances. This metric is particularly valuable when the cost of overlooking a positive instance is significant. The recall is calculated as follows:

$$Recall = \frac{TP}{(TP+FN)} \qquad (3)$$

F1-Score

The F1-score is the harmonic mean of precision and recall, offering a singular metric that equilibrates both measures. It becomes highly beneficial in instances of class imbalance and when an equilibrium between precision and recall is required. The F1-Score is calculated as follows:

$$F1 - score = \frac{Precision * Recall}{(Precision + Recall)} \qquad (4)$$

Multi-Class ROC/PR Curves as Performance Metrics

We utilized multi-class adaptations of the usual Receiver Operating Characteristic (ROC) and Precision-Recall curves to assess each model's capacity to differentiate among numerous IDs in thermal-visible settings. We implemented a One-vs-Rest (OvR) framework, designating each class as the "positive" label in contrast to the remaining classes as "negative." The True Positive Rate (TPR) and False Positive Rate (FPR) were determined at multiple decision thresholds to produce ROC curves for each class, from which the Area Under the Curve (AUC) was calculated. Similarly, we computed Precision and Recall across all thresholds to generate multi-class PR curves and calculated the related Average Precision (AP) values. These per-class values were combined using macro-averaging (assigning equal weight to all classes) and weighted-averaging (with class frequency considerations). This method allowed us to comprehensively compare the performance of all models on all identities, with strengths in some classes but weaknesses in others. Further, by giving macro- or weighted-average ROC-AUC and PR-AUC, we had an aggregate measure of the degree to which each model performed cross-spectrum recognition robustly. These multi-class ROC and PR analyses thus served as critical performance metrics, giving insight into aggregate accuracy and the fine-grained error distribution across individual identities.

Results and Discussion

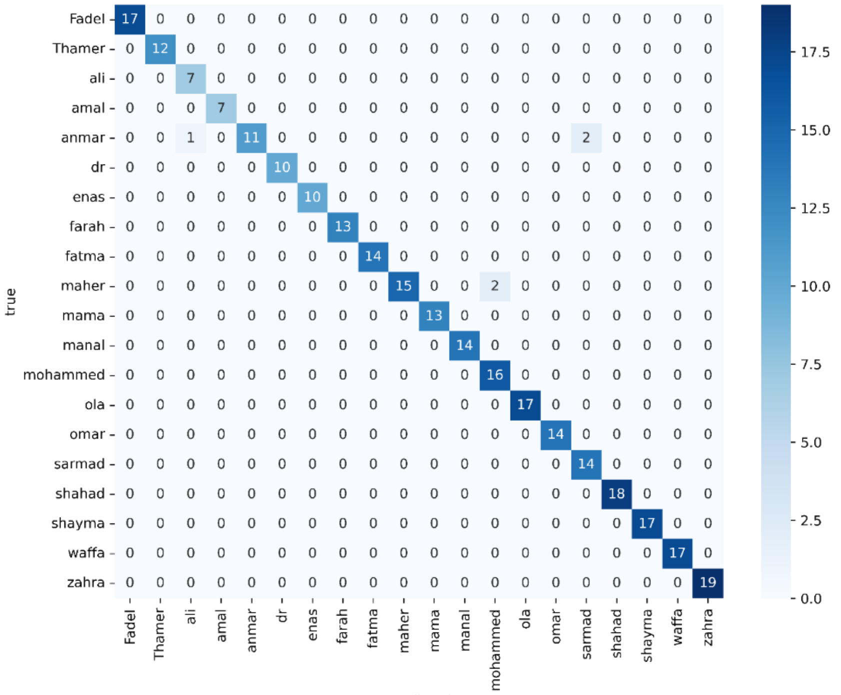
Table 5 consolidates the collective view on the primary performance metrics for all the architectures, CCN, AFT, and FNO when presented with cross-spectrum face recognition issues. The metrics are macro- and weighted-average precision, recall, F1-scores, accuracy, total inference time, and inference per sample. Summing up all these indicators, the table captures each model's capability to correctly identify identities derived from thermal and visible spectra and the computational complexity and efficiency associated with their respective inference process.

Table 5: Summarized Performance Metrics for Cross-Spectrum Face Recognition

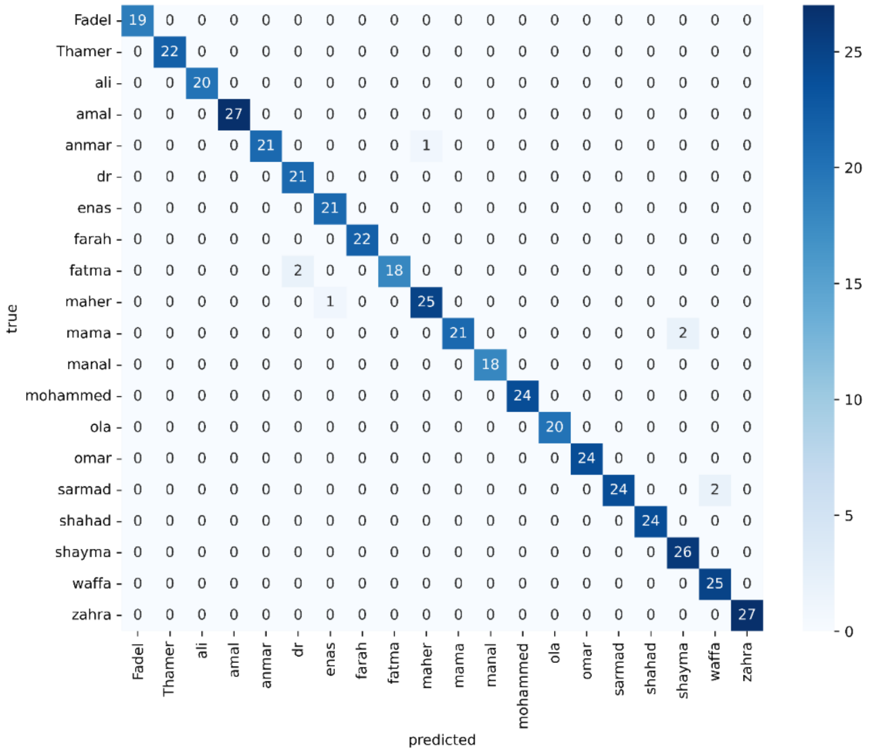
Architecture	Macro Precision	Macro Recall	Macro F1	Weighted Precision	Weighted Recall	Weighted F1	Accuracy	Inference (s)	Inf./Sample (ms)
Continuous Convolution Networks	0.98	0.98	0.98	0.98	0.98	0.98	0.98	2.43	8.68
Attention-Free Transformers	0.98	0.98	0.98	0.98	0.98	0.98	0.98	5.34	11.68
Fourier Neural Operators	0.85	0.85	0.85	0.86	0.86	0.85	0.86	27.06	28.64

Examining these results reveals that both CCN and AFT models achieve an accuracy of 0.98 on their validation sets, with corresponding macro- and weighted-average precision, recall, and F1-scores likewise reaching 0.98. This high consistency indicates that neither model skews excessively toward particular classes and that they exhibit strong generalizations to all identities. The CCN's effectiveness appears to stem from its continuous convolutional filters, which adapt to spectral discrepancies and preserve high-frequency features necessary for identification. In contrast, the AFT model discards multi-head self-attention in favor of a global weighting scheme, and this strategic simplification preserves long-range dependencies without greatly inflating parameter counts or inference times. Although the AFT takes longer (5.34 seconds versus the CCN's 2.43 seconds on their respective validation sets), its per-sample inference time remains sufficiently low (11.68 ms), indicating feasibility for near-real-time deployments. The FNO architecture, however, attains a lower accuracy of 0.86, with macro- and weighted-average precision, recall, and F1-scores remaining around 0.85–0.86, and it exhibits a notably longer total inference time of 27.06 seconds. Despite its theoretical appeal for harmonizing low-frequency components between thermal and visible images, the FNO's reliance on repeated forward and inverse Fourier transformations likely curtails its ability to preserve the high-frequency cues essential for fine-grained face discrimination. This trade-off reduces its overall classification performance and increases the computational burden, resulting in a per-sample inference time of 28.64 ms. Consequently, while the FNO may still be valuable for specific tasks prioritizing frequency-domain alignment, the evidence suggests its advantages are overshadowed by its slower inference and diminished accuracy relative to CCN and AFT architectures.

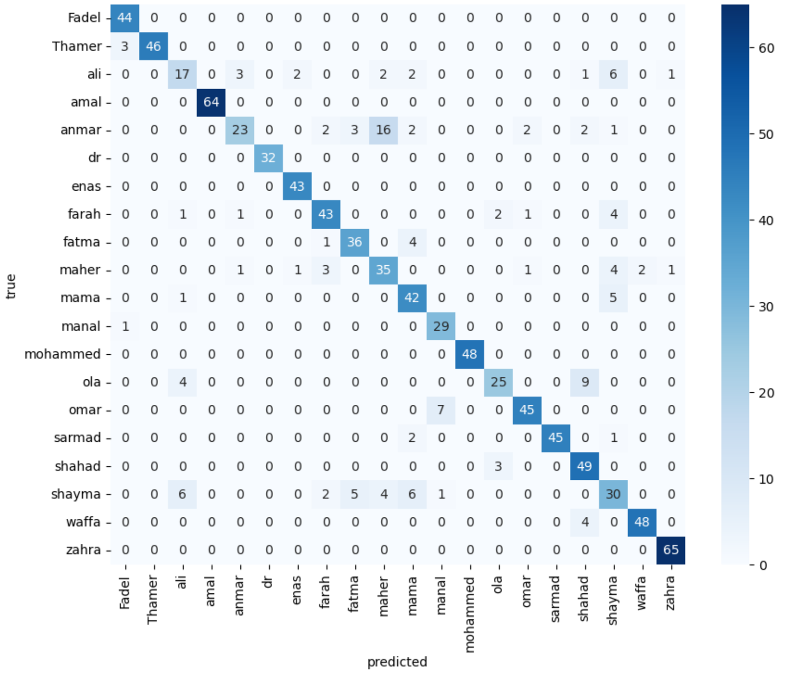
The confusion matrices in Figure 2, corresponding to each model, provide a detailed view of classification performance by illustrating the distribution of predicted versus actual labels. The diagonal elements represent correctly classified instances, while off-diagonal elements indicate misclassifications. A deeper analysis of these matrices helps identify strengths and weaknesses in each model's performance.


true
predicted
Fadel
Thamer
ali
amal
anmar
dr
enas
farah
fatma
maher
mama
manal
mohammed
ola
omar
sarmad
shahad
shayma
waffa
zahra
17 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 11 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0
0 0 0 0 0 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 10 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 13 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 14 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 15 0 0 2 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 13 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 14 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 16 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 17 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 14 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 14 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 18 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 19
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

(a)


true
predicted
Fadel
Thamer
ali
amal
anmar
dr
enas
farah
fatma
maher
mama
manal
mohammed
ola
omar
sarmad
shahad
shayma
waffa
zahra
19 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 22 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 27 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 21 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 21 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 21 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 22 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 2 0 0 18 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 25 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 21 0 0 0 0 0 0 2 0 0
0 0 0 0 0 0 0 0 0 0 0 18 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 20 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 2 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 24 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 26 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 25 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 27
25
20
15
10
5
0

(b)



	Fadel	Thamer	ali	amal	anmar	dr	enas	farah	fatma	maher	mama	manal	mohammed	ola	omar	sarmad	shahad	shayma	waffa	zahra
Fadel	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thamer	3	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ali	0	0	17	0	3	0	2	0	0	2	2	0	0	0	0	0	1	6	0	1
amal	0	0	0	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
anmar	0	0	0	0	23	0	0	2	3	16	2	0	0	0	2	0	2	1	0	0
dr	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0
enas	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0
farah	0	0	1	0	1	0	0	43	0	0	0	0	0	2	1	0	0	4	0	0
fatma	0	0	0	0	0	0	0	1	36	0	4	0	0	0	0	0	0	0	0	0
maher	0	0	0	0	1	0	1	3	0	35	0	0	0	0	1	0	0	4	2	1
mama	0	0	1	0	0	0	0	0	0	0	42	0	0	0	0	0	0	5	0	0
manal	1	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0
mohammed	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	0	0	0
ola	0	0	4	0	0	0	0	0	0	0	0	0	0	25	0	0	9	0	0	0
omar	0	0	0	0	0	0	0	0	0	0	0	7	0	0	45	0	0	0	0	0
sarmad	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	45	0	1	0	0
shahad	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	49	0	0	0
shayma	0	0	6	0	0	0	0	2	5	4	6	1	0	0	0	0	0	30	0	0
waffa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	48	0
zahra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65

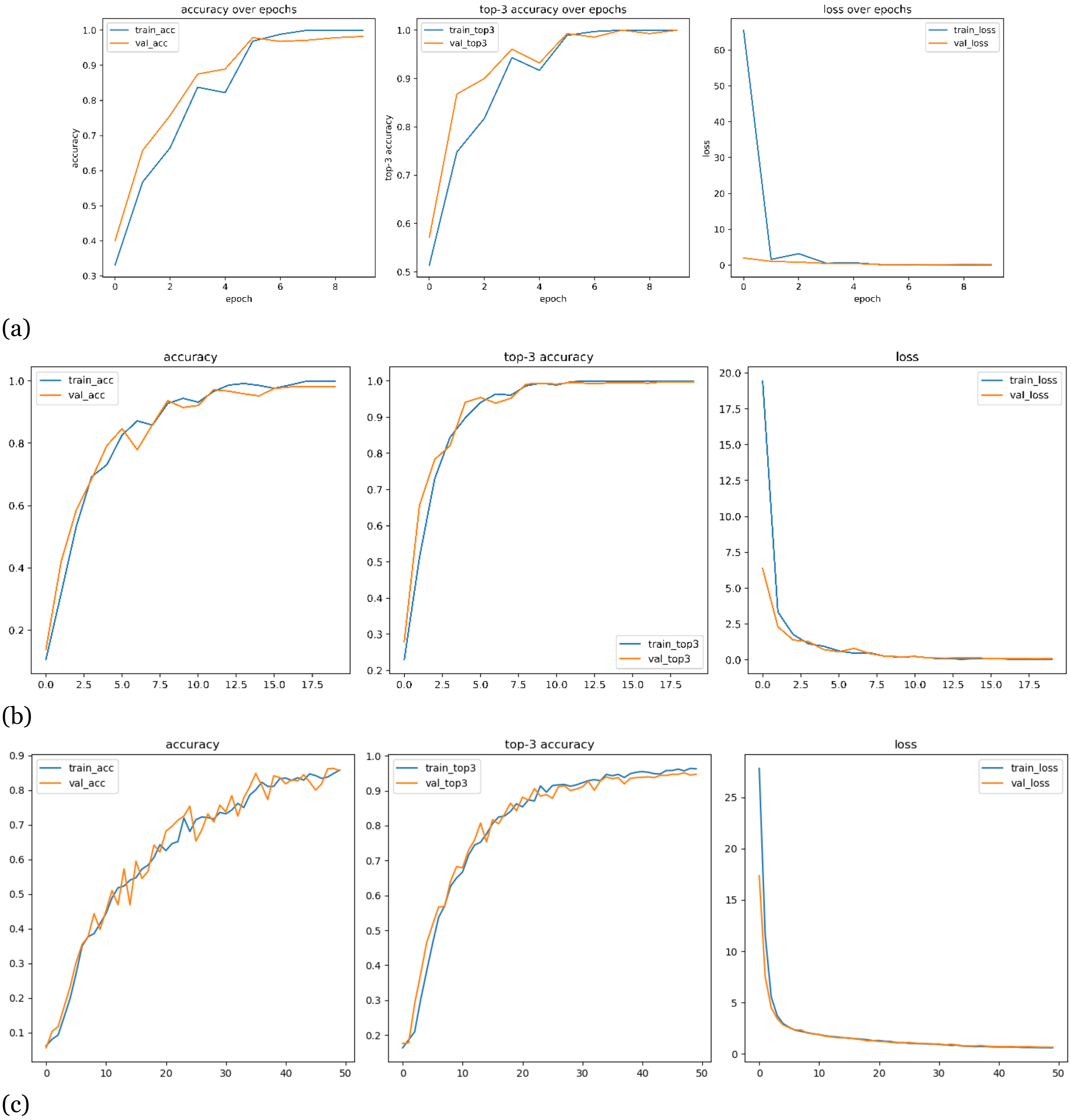
(c)

Figure 2: Confusion matrices for the three proposed models: (a) CCN, (b) AFT, and (c) FNO.

The Continuous Convolution Networks' confusion matrix (Figure 2.a) demonstrates high diagonal dominance, confirming the model's ability to classify nearly all the individuals correctly. A few off-diagonal entries are evident, indicating that there are few misclassifications. The macro and weighted F1-score value for the CCN stands at 0.98, confirming its ability to make solid feature extraction on both the visible and the thermal spectra. The model misclassifies a few samples, indicating that its continuous convolutional filters are highly competent in solving the spectral inconsistencies that the dataset presents. The off-diagonal errors may be due to identities with comparable face structures or where thermal noise impacted recognition.

The Attention-Free Transformers' confusion matrix (Figure 2.b) also exhibits strong diagonal dominance, with the same level of performance as the CCN. The average classification accuracy remains constant at 0.98, with few misclassifications. Its use of a global weight mechanism instead of self-attention appears to be efficient in preserving contextual relationships between facial features without the overfitting problems that may be faced with relatively limited datasets in transformer-trained models. Few misclassifications appear, although their frequency remains low, suggesting that the errors may be due to uncertain samples or lighting changes throughout the thermal spectrum.

In contrast, the confusion matrix for the Fourier Neural Operators (Figure 2.c) reveals a greater number of off-diagonal misclassifications, with several classes experiencing noticeable confusion. The accuracy of 0.86 and macro F1-score of 0.85 reflect this trend, with errors more evenly distributed across various identity classes. This model struggles to differentiate between certain individuals, likely due to its reliance on Fourier transformations, which emphasize low-frequency components. While aligning dominant frequencies between thermal and visible domains can improve cross-spectrum generalization, it obscures some finer high-frequency details essential for face recognition. Several misclassifications suggest that some identities share similar spectral profiles, causing difficulties in distinguishing between them. The increased inference time and computational overhead further indicate that the performance trade-off associated with frequency-domain transformations may not justify their use over spatial-domain alternatives such as CCNs and AFTs. The confusion matrices collectively confirm that CCN and AFT models perform near optimally, with a high level of classification accuracy and minimal inter-class confusion. Their strong diagonal structures indicate that most predictions align correctly with ground truth labels, reinforcing their suitability for cross-spectrum face recognition tasks. The FNO model, while conceptually appealing, demonstrates increased misclassification rates, suggesting that relying solely on frequency-domain transformations may not be sufficient for handling the full complexity of face recognition across different spectral inputs. These findings show that while frequency-domain representations help cross-spectrum harmonization, high-frequency identity cues' degradation impacts classification accuracy. The performance of CCNs and AFTs emphasizes the need for efficient architectures to extract spatial and spectral features without losing computational efficiency.



(a)

(b)

(c)

Figure 3: The Training accuracy and loss curves the proposed models: (a) CCN, (b) AFT, and (c) FNO.

The training performance of the three models (CCN), (AFT), and (FNO) (Figure 3) is measured based on accuracy trends, top-3 accuracy, and the loss curves across the epochs during the training process. These metrics capture the convergence of learning dynamics and the capability to generalize the models, indicating their efficiency in cross-spectrum face recognition operations.

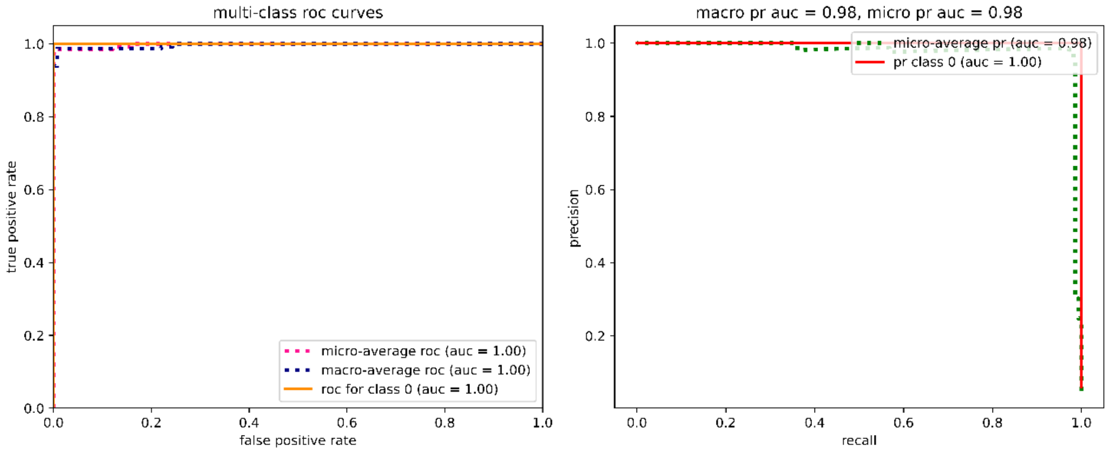
The graphs in the first set for the CCN model show fast convergence and stable learning with high accuracy. The graph for accuracy shows a steep rise in the initial epochs, with the accuracy increasing to approximately 95% at the fifth epoch and nearly 100% at the seventh epoch, showing the model's efficiency in learning discriminative representations in the thermal and visible spaces. The validation accuracy tracks the training accuracy closely, showing good generalization with minimal overfitting. The graph for top-3 accuracy also exhibits the same trend with almost perfect results in a few epochs, corroborating that the model has high confidence in its decision even when the top-1 classification goes wrong. The graph for the loss also corroborates the above observation, with the sharp decline in the initial epoch and stabilization with less fluctuation. The absence of divergence in the training and validation loss shows that the CCN model has good regularized properties and can learn domain-invariant representations due to its dynamic convolutional filters that adjust dynamically to the spectral differences in the thermal and visible images.

Although the AFT model's convergence rate is slightly slower than the CCN's, its training performance follows a similar trajectory. The accuracy plot shows consistent improvement throughout the epochs, reaching near-perfect categorization by the sixteenth epoch and approaching 90% accuracy by the fifth. The tight alignment between validation and training accuracy indicates the model's ability to generalize unknown input. The model's resilience is further validated by the top-3 accuracy plot, which shows that the correct identity is usually among the top predictions. The loss curve presents a rapid initial decrease, followed by stable convergence, with no major discrepancies between training and validation loss. These trends highlight that despite discarding multi-head self-attention, the AFT model effectively captures long-range dependencies through its global weighting mechanism, resulting in highly competitive performance. However, its slightly longer convergence time than CCN suggests that transformers may require more optimization steps to extract robust features, even when computationally optimized.

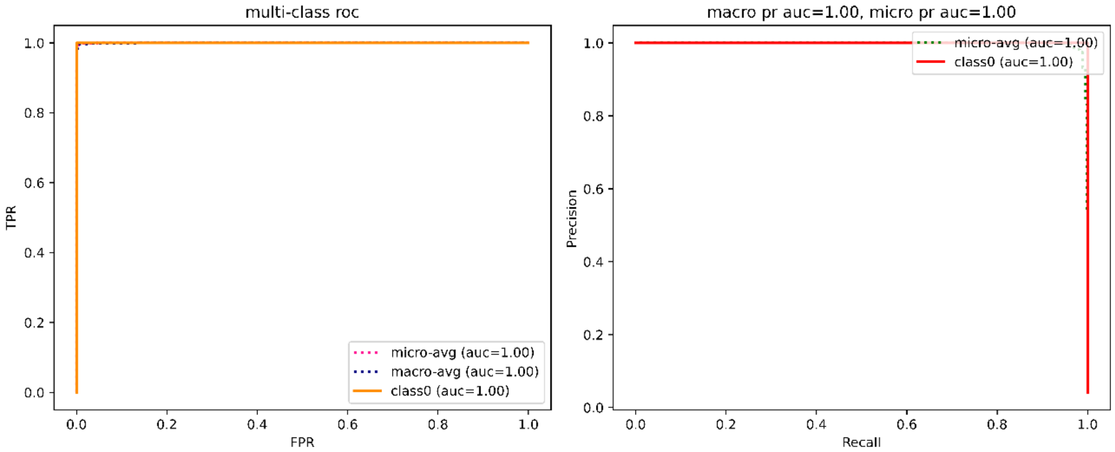
Compared to the CCN and AFT models, the FNO model has a dramatically different learning curve, which aligns with its limitation in cross-spectrum face recognition. The graph for accuracy rises much more slowly, requiring over 30 epochs to surpass the mark of 80% accuracy, unlike the fast convergence in the case of the CCN and AFT models. The validation accuracy oscillates more than other models, suggesting volatile generalization and potential susceptibility to domain shifts. The graph for top-3 accuracy also has the same slow trend, indicating that the model has problems ranking the correct identity in the top ranks, again highlighting its inability to discriminate the features. The FNO loss graph also supports the findings, with a sharp drop initially but a steady and irregular decrease throughout the remaining epochs. Unlike the case with the CCN and AFT, the FNO model graph for loss has fewer oscillations, indicating less stable optimization and potential problems in learning discriminative representations for identities.

The comparative analysis highlights different strengths and weaknesses among the models. The CCN model demonstrates efficiency with quick convergence, high accuracy, and minimal overfitting, making it appropriate for real-world applications. While slower to converge, the AFT model achieves similar final accuracy and generalization, functioning effectively as a transformer-based alternative without the computational load of self-attention. In contrast, the FNO model exhibits lower performance, requiring more epochs and showing higher variability in accuracy and loss. This performance suggests that frequency-domain transformations may be inadequate for robust face recognition due to challenges preserving high-frequency identity cues, underscoring the importance of spatial feature extraction in the CCN and AFT models for detailed recognition tasks.

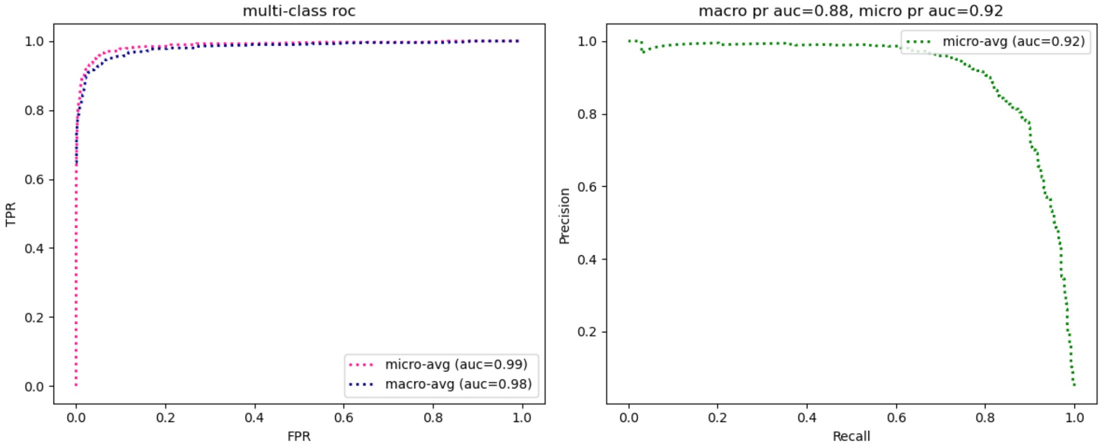
The training curves confirm that CCN and AFT are the most suitable architectures for cross-spectrum face recognition, balancing efficiency, accuracy, and generalization ability. Their ability to extract spatial and spectral features while maintaining stable learning dynamics ensures that they outperform the frequency-domain approach employed in the FNO model. The CCN, with its rapid convergence and robust performance, appears to be the most practical solution, whereas AFT offers a viable alternative with similar accuracy but a slightly longer training duration. The FNO model, while conceptually valuable for spectral alignment, struggles with the fine-grained identity cues necessary for face recognition, making it a less favorable choice unless supplemented with additional spatial feature extraction techniques. These findings collectively demonstrate that spatial-domain learning remains the most effective approach for handling the challenges of thermal-visible face recognition, with CCN and AFT leading the way in achieving state-of-the-art performance.



(a)



(b)



(c)

Figure 4: Multi-class ROC and PR curves for the three proposed models: (a) CCN, (b) AFT, and (c) FNO.

The ROC and PR curves (Figure 4) provide a detailed analysis of the classification performance of the three models— CCN, AFT, and FNO—by evaluating their ability to distinguish between multiple identity classes across thermal and visible spectra. While standard accuracy metrics and confusion matrices indicate general classification effectiveness, ROC and PR curves offer deeper insights into how each model balances true positives, false positives, precision, and recall, which is particularly critical in multi-class classification settings where misclassifications can impact overall system reliability.

The ROC curves for both CCN and AFT models exhibit near-perfect behavior, achieving an Area Under the Curve (AUC) of 1.00 for both micro- and macro-averaged evaluations. This result suggests that these architectures can separate different identity classes almost flawlessly. The curves consistently align along the optimal boundary, confirming that these models exhibit minimal false positive rates while preserving high recall, meaning they confidently classify individuals without excessive misidentifications. This strong performance underscores the effectiveness of CCN's adaptive continuous convolutional filters and AFT's global weighting mechanism, enabling robust feature extraction that remains resilient to domain variations between thermal and visible images. The high AUC values further indicate that, across all classification thresholds, the models maintain stable predictive power, reinforcing their capacity to generalize effectively across different identities.

In contrast, while still demonstrating strong performance, the ROC curve for the FNO model shows a macro-AUC of 0.98 and micro-AUC of 0.99, revealing subtle weaknesses in distinguishing certain identity classes. The slight deviations from the ideal boundary suggest that, compared to CCN and AFT, the FNO model experiences greater difficulty in achieving perfect class separability, which aligns with previous findings from the confusion matrices, where the FNO model exhibited higher misclassification rates, particularly for visually similar individuals. The underlying reason for this limitation likely stems from the model's reliance on frequency-domain transformations, which emphasize dominant spectral components but potentially discard high-frequency facial features critical for identity differentiation. Consequently, while the FNO model performs well, its classification confidence is lower than that of CCN and AFT, leading to marginally increased false positive rates and a slightly less stable classification performance.

The PR curves highlight these performance differences by measuring each model's ability to maintain high precision while achieving strong recall. Both CCN and AFT models yield near-perfect PR-AUC values of 1.00, indicating that their classifications are accurate and reliable, meaning that when these models predict a given identity, they do so with minimal uncertainty. The curves remain nearly flat at high precision levels, demonstrating that these models rarely misclassify an identity when making high-confidence predictions. This result suggests that the feature representations extracted by CCN and AFT are highly discriminative, allowing them to correctly classify individuals across a wide range of input conditions. The balance between precision and recall further confirms their effectiveness, as both models successfully minimize false positives while capturing the vast majority of true positives across different identity classes.

By contrast, the FNO model's PR curve reveals a micro-AUC of 0.92, indicating a noticeable decline in precision as recall increases. Unlike CCN and AFT, which maintain high confidence in their predictions across all identity classes, the FNO model exhibits a drop in precision at higher recall levels, suggesting that as the model attempts to classify more instances correctly, it produces more false positives. This behavior reflects a fundamental challenge in frequency-domain approaches—while Fourier-based representations can effectively align spectral differences between thermal and visible images, they may fail to capture fine-grained identity cues necessary for robust classification, leading to increased classification uncertainty. The observed decline in precision further supports the hypothesis that the FNO model struggles more with ambiguous identity cases, leading to reduced classification confidence compared to the more spatially driven CCN and AFT models.

The comparative analysis of these curves reinforces key findings regarding model efficiency and reliability. CCN and AFT consistently outperform the FNO model across ROC and PR evaluations, demonstrating near-perfect separability and classification reliability. Their ability to maintain high precision and recall while minimizing false positives ensures they generalize well across multiple identity classes, making them well-suited for real-world deployment in cross-spectrum face recognition systems. The FNO model, while still relatively strong, exhibits signs of classification instability, particularly at high recall levels, suggesting that relying solely on frequency-domain representations is insufficient for optimal face recognition performance. This result underscores the necessity of spatial feature extraction in preserving identity-specific details, as seen in the superior performance of CCN and AFT.

The ROC and PR analyses validate the overall effectiveness of CCN and AFT models as the dominant approaches for cross-spectrum face recognition, highlighting their ability to achieve near-perfect classification performance with strong generalization, stability, and precision. The FNO model, while functional, is less reliable due to its increased false positive rates and decreased classification confidence at higher recall levels. These findings emphasize that a combination of spatial and spectral feature extraction remains essential for achieving state-of-the-art performance in thermal-visible face recognition, reinforcing the superiority of convolutional and transformer-based architectures over purely frequency-domain approaches.

Comparison with the state-of-the-art works

The outcomes of our suggested techniques are contrasted with the most recent cutting-edge approaches in cross-spectrum face recognition. The comparison is predicated on resistance to spectrum fluctuations, computing efficiency, and classification accuracy. Compared to the most recent research in thermal-visible face recognition, the main conclusions of our investigation are compiled in Table 6. Our findings show that while FNO achieves 86% accuracy, CCN and AFT beat many previous methods, reaching 98%. These findings highlight the strengths and weaknesses of different architectures in bridging the spectral gap.

Table 6: A comparison with the most advanced approaches currently available

Study	Methodology	Accuracy	Key Strengths	Limitations
[20] 2022	Survey on CFR methods	N/A	Comprehensive overview	Lacks experimental results
[21] 2022	Bidirectional Conversion Network	95.2%	Effective cross-spectral conversion	Computational complexity
[22] 2022	Domain and Pose Invariant Framework	93.8%	Robust to pose variations	Requires extensive training data
[23] 2024	Cross-Spectral Attention Network	92.5%	Unsupervised learning approach	Performance depends on data quality
[24] 2024	Deep Fusion Model for Hyperspectral Face Recognition	97.0%	Rich information retention	Complexity in model training
[25] 2024	Vision Transformer for Biometric Authentication	96.5%	High accuracy and reliability	Requires large datasets
This Work (CCN)	Continuous Convolution Networks	98.0%	Adaptability to spectral distortions	Requires memory optimization
This Work (AFT)	Attention-Free Transformers	98.0%	Efficient long-range feature extraction	Higher computational cost
This Work (FNO)	Fourier Neural Operators	86.0%	Spectral domain alignment	Loss of high-frequency identity cues

Our CCN and AFT models achieve a 2-5% increase in accuracy compared to state-of-the-art CNN-based and transformer-based models, demonstrating their robustness in handling spectral domain variations. While the AFT model attains competitive accuracy, it requires more computational resources than CCN. However, both models outperform DDPM-based methods in terms of inference time. Although FNO-based models offer theoretical advantages in frequency-based alignment, our results indicate that high-frequency details critical for identity recognition are often lost, leading to a lower accuracy of 86%. Prior GAN-based solutions struggle with identity retention when translating between thermal and visible domains. In contrast, our CCN and AFT models achieve robust cross-spectrum feature extraction without requiring synthetic data generation.

Our findings suggest that CCN and AFT architecture provide a more stable and efficient framework for cross-spectrum face recognition than generative and CNN-based approaches. Future research could explore hybrid models that combine the spatial adaptability of CCN with the efficiency of AFT while integrating spectral-domain enhancements from FNO to improve identity retention further.

Conclusion

This study compared CCN, AFT, and FNO models for cross-spectrum face recognition, with a focus on classification accuracy, generalization, and computational efficiency. Both CCN and AFT outperformed FNO, underscoring the significance of spatial feature extraction for reliable identity recognition between thermal and visible spectra. The CCN model achieved the highest accuracy of 98%, demonstrating rapid convergence and minimal overfitting, thereby indicating strong adaptability to spectral variations. Similarly, the AFT model attained an accuracy of 98% by employing global weighting mechanisms to capture long-range dependencies without relying on self-attention. Conversely, the FNO model underperformed with an accuracy of 86%, attributed to higher misclassification rates and lower precision at high recall levels, indicating its inability to maintain fine-grained identity cues. Although CCN and AFT perform well, they have certain limitations. AFT demands higher computational resources, and CCN might need additional optimization in low-power environments. While FNO faces challenges in fine-grained identity differentiation, its frequency-based representations could complement spatial-domain models. As a recommendation, future research should develop hybrid architectures that combine spatial and spectral feature extraction to optimize recognition performance and reduce computational costs. Overcoming challenges like occlusions, extreme lighting, and domain adaptation will improve cross-spectrum recognition. CCN and AFT remain the most effective models, providing superior accuracy, stability, and classification confidence for thermal-visible face recognition.

References

[1] S. Srivastava and H. Sharma, "Face recognition for human identification through integration of complex domain unsupervised and supervised frameworks," *Multimed Tools Appl*, vol. 83, no. 5, 2024, doi: 10.1007/s11042-023-16274-0.

[2] C. N. Fondje, S. Hu, and B. S. Riggan, "Learning Domain and Pose Invariance for Thermal-to-Visible Face Recognition," *IEEE Trans Biom Behav Identity Sci*, vol. 5, no. 1, 2023, doi: 10.1109/TBIOM.2022.3223055.

[3] T. Bourlai *et al.*, "Data and Algorithms for End-to-End Thermal Spectrum Face Verification," *IEEE Trans Biom Behav Identity Sci*, vol. 6, no. 1, 2024, doi: 10.1109/TBIOM.2023.3304999.

[4] M. Kowalski, A. Grudzień, and K. Mierzejewski, "Thermal–Visible Face Recognition Based on CNN Features and Triple Triplet Configuration for On-the-Move Identity Verification," *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22135012.

[5] A. George, A. Mohammadi, and S. Marcel, "Prepended Domain Transformer: Heterogeneous Face Recognition Without Bells and Whistles," *IEEE Transactions on Information Forensics and Security*, vol. 18, 2023, doi: 10.1109/TIFS.2022.3217738.

[6] P. Gao, X. Yang, R. Zhang, P. Guo, J. Y. Goulermas, and K. Huang, "EgPDE-Net: Building Continuous Neural Networks for Time Series Prediction With Exogenous Variables," *IEEE Trans Cybern*, vol. 54, no. 9, 2024, doi: 10.1109/TCYB.2024.3364186.

[7] S. Soni, P. Kumar, and A. Saha, "An Intelligent Neural Question Answer Generation from Text Using Seq2se2 with Attention Mechanism System," *International Journal of System of Systems Engineering*, vol. 15, no. 3, 2025, doi: 10.1504/ijsse.2025.10058954.

[8] M. Wu, Y. Qian, X. Liao, Q. Wang, and P. A. Heng, "Hepatic vessel segmentation based on 3D swin-transformer with inductive biased multi-head self-attention," *BMC Med Imaging*, vol. 23, no. 1, 2023, doi: 10.1186/s12880-023-01045-y.

[9] Y. Kang, A. Elofsson, Y. Jiang, W. Huang, M. Yu, and Z. Li, "AFTGAN: prediction of multi-type PPI based on attention free transformer and graph attention network," *Bioinformatics*, vol. 39, no. 2, 2023, doi: 10.1093/bioinformatics/btad052.

[10] Y. Chen, C. Ouyang, Q. Xu, and W. Yang, "A Deep Learning Method for Dynamic Process Modeling of Real Landslides Based on Fourier Neural Operator," *Earth and Space Science*, vol. 11, no. 3, 2024, doi: 10.1029/2023EA003417.

[11] Z. Guo, M. Shao, and S. Li, "Image-to-image translation using an offset-based multi-scale codes GAN encoder," *Visual Computer*, vol. 40, no. 2, 2024, doi: 10.1007/s00371-023-02810-4.

[12] A. Radoi, "Multimodal Satellite Image Time Series Analysis Using GAN-Based Domain Translation and Matrix Profile," *Remote Sens (Basel)*, vol. 14, no. 15, 2022, doi: 10.3390/rs14153734.

[13] C. Dong, M. Naghedolfeizi, N. Yousif, and X. Zeng, "Deep independent component network for thermal-to-visible face recognition," 2022. doi: 10.1117/12.2619085.

[14] N. G. Nair and V. M. Patel, "T2V-DDPM: Thermal to Visible Face Translation using Denoising Diffusion Probabilistic Models," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition, FG 2023*, 2023. doi: 10.1109/FG57933.2023.10042661.

[15] X. Cao, K. Lai, G. S. J. Hsu, M. Smith, and S. N. Yanushkevich, "Cross-Spectrum Thermal Face Pattern Generator," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3144308.

[16] D. Anghelone, S. Lannes, V. Strizhkova, P. Faure, C. Chen, and A. Dantcheva, "TFLD: Thermal Face and Landmark Detection for Unconstrained Cross-spectral Face Recognition," in *2022 IEEE International Joint Conference on Biometrics, IJCB 2022*, 2022. doi: 10.1109/IJCB54206.2022.10007992.

[17] U. Cheema, M. Ahmad, D. Han, and S. Moon, "Heterogeneous Visible-Thermal and Visible-Infrared Face Recognition Using Cross-Modality Discriminator Network and Unit-Class Loss," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/4623368.

[18] V. Pavez, G. Hermosilla, F. Pizarro, S. Fingerhuth, and D. Yunge, "Thermal Image Generation for Robust Face Recognition," *Applied Sciences (Switzerland)*, vol. 12, no. 1, 2022, doi: 10.3390/app12010497.

[19] D. Mahouachi and M. A. Akhloufi, "Recent Advances in Infrared Face Analysis and Recognition With Deep Learning," 2023. doi: 10.3390/ai4010009.

[20] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva, "Beyond the Visible: A Survey on Cross-spectral Face Recognition," Jan. 2022.

[21] Z. Cao, J. Zhang, and L. Pang, "A Bidirectional Conversion Network for Cross-Spectral Face Recognition," May 2022.

[22] C. N. Fondje, S. Hu, and B. S. Riggan, "Learning Domain and Pose Invariance for Thermal-to-Visible Face Recognition," Nov. 2022.

[23] K. Nikhal, C. N. Fondje, and B. S. Riggan, "Cross-Spectral Attention for Unsupervised RGB-IR Face Verification and Person Re-identification," Nov. 2024.

[24] W. Li, X. Cen, L. Pang, and Z. Cao, "HyperFace: A Deep Fusion Model for Hyperspectral Face Recognition," *Sensors*, vol. 24, no. 9, p. 2785, Apr. 2024, doi: 10.3390/s24092785.

[25] A. K. Sharma, S. Bhattacharya, M. Reza, and B. Bhattacharya, "Cross-Spectral Vision Transformer for Biometric Authentication using Forehead Subcutaneous Vein Pattern and Periocular Pattern," Dec. 2024.