1. Introduction

Gender classification has become a fundamental task in computer vision, with numerous applications across various domains including in-cabin driver monitoring systems, human-computer interaction, video surveillance, retail analytics, and psychological analysis. Traditionally, researchers have focused on gender classification using visible spectrum images of the human face. However, the performance of these systems can be significantly affected by challenging environmental factors such as varying illumination conditions, shadows, occlusions, and the time of day.

To overcome these limitations, there has been a growing interest in exploring alternative or complementary sensing modalities, such as **thermal imaging**. Thermal imaging offers several advantages as it does not rely on external illumination and provides a distinct perspective on an imaged scene compared to conventional visible light sensors. This makes it a potentially more robust solution for gender classification in diverse and uncontrolled environments. Furthermore, thermal imaging can easily detect people even in total darkness, expanding its applicability in security systems. Beyond security, thermal signatures can provide complementary information in human-computer interaction, potentially revealing subtle physiological indicators relevant to gender.

Despite the benefits, thermal images typically lack some of the detailed facial definitions present in visible spectrum images, posing a challenge for accurate classification. To address this, the application of **deep learning**, particularly **Convolutional Neural Networks (CNNs)**, has shown significant promise in learning intricate patterns from thermal data for gender classification.

This paper investigates the effectiveness of deep learning models for gender detection using thermal facial images. We utilize two publicly available thermal image datasets, the **Tufts University Thermal Face dataset** and the **Charlotte-ThermalFace dataset**, both individually and in combination, to train and evaluate a range of state-of-the-art CNN architectures including **AlexNet**, **VGG**, **InceptionV3**, **ResNet50**, and **EfficientNet**. We address the differences in channel availability between the datasets and enhance the data through **image augmentation** techniques. Furthermore, we tackle the class imbalance present in the Tufts dataset to ensure robust training. To further advance the field, we propose a **novel CNN architecture** based on the ResNet framework, incorporating a **channel input adapter** to handle varying input channels and **Squeeze and Excitation (SE) blocks** within its layers to enhance feature discrimination, along with a tailored final classifier.

The primary contributions of this paper include:

- A comprehensive evaluation of several state-of-the-art CNN models for gender classification on thermal facial images using the Tufts and Charlotte datasets.

- An investigation into the impact of combining datasets with differing channel characteristics.
- The development and evaluation of a novel CNN architecture specifically designed for thermal image-based gender detection, incorporating channel adaptation and attention mechanisms.
- An analysis of the challenges and potential of deep learning for gender classification using thermal imaging.

The remainder of this paper is structured as follows: Section 2 provides a review of related work in gender classification using both traditional and deep learning methods with visible, near-infrared, and thermal imagery. Section 3 details the datasets used and the methodology employed, including preprocessing, augmentation techniques, and the architecture of the proposed CNN model. Section 4 presents the experimental results and a comparative analysis of the different models. Section 5 discusses the implications and limitations of our findings, and Section 6 concludes the paper with potential directions for future research.

2. Literature Review

The task of gender classification has been extensively studied in computer vision. Early approaches often relied on **conventional machine learning methods** and feature extraction techniques applied to visible spectrum images. Makinen and Raisamo and Reid et al. provided detailed surveys of these methods. Initial techniques involved training neural systems on small sets of frontal face images. Later, methods incorporated 3D head structure and image intensities for gender characterization. **Support Vector Machines (SVMs)** were also widely used, demonstrating competitive performance compared to other traditional classifiers. Techniques like AdaBoost, utilizing low-resolution grayscale images, and methods addressing perspective invariant recognition were also explored. More recently, researchers utilized local image descriptors like the Webers Local Surface Descriptor (WLD) and features based on shape, texture, and color extracted from frontal faces, achieving high accuracy on benchmark datasets like FERET.

Recognizing the limitations of visible spectrum-based methods, researchers began to explore the potential of deducing gender information from other modalities, including **thermal and Near-Infrared (NIR) spectra**. Chen and Ross are noted as early proponents of human face-based gender classification systems using both thermal and NIR data, employing conventional feature extraction methods and classifiers like SVM, LDA, Adaboost, random forest, Gaussian mixture models, and multi-layer perceptrons. Their findings suggested that SVM with histogram-based gender classification yielded better performance on NIR and thermal spectra. Nguyen and Park proposed a gender classification system using joint visible and thermal spectrum data of the human body, utilizing feature extractors like Histogram of Oriented Gradients (HoG) and Multi-Local Binary Patterns (MLBP). Their results indicated improved accuracy by combining data from both modalities. Similarly, Abouelenien et al. explored multimodal datasets including audiovisual, thermal, and physiological

recordings for automatic gender classification, again relying on conventional machine learning algorithms.

The advent of **deep learning** and the success of **CNNs** in various computer vision tasks, particularly where high accuracy and robustness are required, led to their application in gender classification. Canziani et al. listed numerous pretrained models suitable for various applications. Dwivedi and Singh provided a comprehensive review of deep learning methodologies for robust gender classification using visible spectrum datasets. Ozbulak et al. investigated fine-tuning and SVM classification using CNN features for age and gender classification on visible datasets, demonstrating that transferred models can outperform task-specific models. Manyala et al. explored CNN-based methods for gender classification using NIR periocular images, achieving promising results. Baek et al. used combined visible and NIR data with two CNN architectures for robust gender classification from full human body images in surveillance environments.

In the domain of **thermal image-based gender classification**, Farooq et al. conducted a comprehensive performance estimation of state-of-the-art CNNs, including AlexNet, VGG-19, MobileNet-V2, Inception-V3, ResNet-50, ResNet-101, DenseNet-121, DenseNet-201, and EfficientNet-B4, using the **Tufts thermal faces dataset** and the **Carl thermal faces dataset**. They also proposed a new CNN architecture, **GENNet**, specifically for this task. Li et al. focused on detecting age and gender from thermal images for personal thermal comfort prediction, utilizing a newly established dataset of thermal and visible-light images. They evaluated the performance of ResNet-50, ResNet-101, EfficientNet, and Inception v3, finding ResNet-50 to achieve a high gender accuracy on their thermal dataset. Chatterjee and Zaman proposed a deep-learning approach for general thermal image classification, utilizing pretrained ResNet-50 and VGGNet-19 and exploring the impact of Kalman filtering for denoising on the **Tufts** and **Charlotte-ThermalFace datasets**. Keerthi et al. investigated gender classification optimization with thermal images using InceptionV3 and AlexNet, also utilizing the "tufts" dataset.

These studies highlight the growing interest and potential of using deep learning techniques for gender classification based on thermal imagery. Our work builds upon this foundation by providing a comparative analysis of several prominent CNN architectures on the **Tufts** and **Charlotte** datasets, addressing the challenges of varying input channels, class imbalance, and further proposing and evaluating a novel architecture tailored for this specific task with the incorporation of channel adaptation and Squeeze-and-Excitation mechanisms. This research aims to contribute to the advancement of robust and accurate gender detection systems using thermal imaging in diverse real-world applications.

3. Methodology

3.1 Datasets

3.1.1 Tufts University Thermal Face Dataset

The Tufts University Thermal Face Dataset represents a comprehensive multimodal collection comprising over 10,000 images across various modalities acquired from a diverse cohort of 113 participants (74 females, 39 males). For our research, we specifically utilized the thermal subset containing approximately 1,400 images. The age distribution spans from 4 to 70 years, with subjects originating from more than 15 countries, thus providing substantial demographic variability. Image acquisition was conducted using a FLIR Vue Pro thermal camera under controlled indoor environmental conditions. Participants were positioned at a standardized distance from the imaging apparatus to maintain consistency. For our investigation, we specifically utilized two subsets: TDIRE (Emotion), which contains images depicting five distinct facial expressions (neutral, smile, eyes closed, shocked, and with sunglasses), and TDIRA (Around), which encompasses images captured from nine different camera positions arranged in a semicircular configuration around each participant. A significant challenge encountered with this dataset was the pronounced gender imbalance, with approximately 30.32% female and 69.68% male images. To mitigate this imbalance and enhance model robustness, we implemented targeted data augmentation techniques specifically for the underrepresented female class, including controlled geometric transformations and intensity adjustments while preserving critical thermal signature characteristics.

(Example Images from tufts dataset)

3.1.2 Charlotte-ThermalFace Dataset

The Charlotte-ThermalFace Dataset comprises approximately 10,364 thermal facial images from 10 subjects, collected under varying conditions (e.g., distance, head position, temperature). This dataset was not specifically created for gender detection tasks, but we repurposed it for our gender classification research. Based on image characteristics, we infer that data acquisition likely employed a FLIR-based thermal imaging system. In contrast to the Tufts collection, the Charlotte dataset exhibits near-perfect gender balance with approximately 50.10% female and 49.90% male. This balanced distribution provided an advantageous counterpoint to the gender imbalance present in the Tufts dataset.

(Example Images from charlotte dataset)

3.1.3 Combined Dataset

To enhance data diversity and expand the training corpus, we constructed a combined dataset by integrating the Tufts and Charlotte collections

following a systematic merging protocol. A significant technical challenge encountered during this integration was the channel discrepancy between datasets—the Charlotte images were originally single-channel thermal representations, whereas the Tufts dataset employed a three-channel format. To address this incompatibility, we implemented channel replication for the Charlotte images, duplicating the single thermal channel across three channels to establish format consistency with the Tufts data structure. Furthermore, to prevent model bias towards the overrepresented class, we carefully balanced the gender distribution by selecting an equal number of images per gender category through strategic sampling. This integration yielded a substantially enlarged dataset of approximately 11,921 images with perfect gender balance (50% female, 50% male), thereby providing our models with enhanced training diversity spanning different thermal imaging conditions, acquisition parameters, and subject characteristics.

In addition to the primary datasets, we designed cross-dataset experimental protocols to rigorously evaluate model generalization capabilities across different thermal imaging domains. These experiments comprised two principal configurations: Tufts-to-Charlotte (training on Tufts data and evaluating on Charlotte) and Charlotte-to-Tufts (training on Charlotte and evaluating on Tufts). This cross-domain validation approach enables assessment of our models' ability to generalize across varying thermal imaging conditions, camera specifications, and data collection protocols—a critical factor for real-world deployment scenarios where thermal imaging parameters may differ substantially from training conditions.

Table 1: Summary of Datasets

Dataset	Size (Images)	Gender Distribution	Channels
Tufts	~1,400	30.32% F, 69.68% M	Three (thermal representation)
Charlotte	~10,000	50.10% F, 49.90% M	One (thermal grayscale)
Combined	~11,921	50.00% F, 50.00% M	Three-channel format

3.2 Data Preprocessing and Augmentation

Our data preprocessing and augmentation pipeline was meticulously designed to address the unique challenges of thermal facial image analysis for gender classification. The pipeline incorporated several carefully considered stages to ensure optimal model performance and generalization.

3.2.1 Dataset Organization and Partitioning

We structured our datasets according to a standardized hierarchical organization to facilitate efficient training and evaluation. Each dataset (Tufts, Charlotte, and Combined) was systematically partitioned into training and testing subsets using a subject-disjoint approach. This critical design choice ensured that images from the same individual never appeared

in both training and testing sets, thus preventing identity-based information leakage that could artificially inflate performance metrics. We implemented an 80:20 train-test split ratio, stratified by gender to maintain proportional representation across partitions.

For the Tufts dataset, we addressed the inherent gender imbalance during partitioning, ensuring that the disproportionate male-to-female ratio was consistently reflected in both training and testing subsets. In the Charlotte dataset, the near-perfect gender balance was preserved throughout the partition process. For the combined dataset, we implemented balanced sampling to achieve gender parity while maintaining subject-level separation between training and testing sets.

Figure 1: Subject-Disjoint Dataset Partitioning Schema - A diagram showing the hierarchical organization and separation of subjects by gender across train/test splits.

3.2.2 Image Normalization and Standardization

Thermal imaging data presents unique challenges due to variations in sensor calibration, environmental conditions, and temperature ranges. To mitigate these issues, we implemented a comprehensive normalization protocol:

All thermal images were normalized using mean-centering with a value of 0.5 and standard deviation scaling of 0.5. This approach was selected over the conventional ImageNet normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) as it proved more effective for thermal imagery during our preliminary experiments, likely due to the fundamentally different intensity distribution characteristics of thermal versus visible spectrum images.

The Charlotte dataset's single-channel thermal images required special handling when used in models designed for multi-channel inputs. For our hybrid models specifically designed for thermal data, we maintained the single-channel representation, utilizing the grayscale transformation to preserve thermal intensity information. For standard RGB-designed architectures, we expanded the single channel through replication to maintain compatibility while preserving the original thermal information.

Images were resized according to model-specific requirements—224×224 pixels for AlexNet, VGG, ResNet, and EfficientNet; 299×299 pixels for Inception. This standardization ensured consistent spatial dimensions while preserving the aspect ratio through center cropping, thus maintaining the integrity of facial thermal patterns.

Figure 2: Thermal Image Normalization Process - Visual comparison showing raw thermal images alongside their normalized and channel-harmonized versions from both Tufts and Charlotte datasets.

Table 2: Model-Specific Normalization Parameters

Model Type	Input Size	Normalization Values	Channels	Rationale
AlexNet/VGG/ ResNet/ EfficientNet	224×224	mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5]	3	Optimized for thermal intensity distribution
Inception	299×299	mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5]	3	Maintained larger input size for finer detail capture
Hybrid Models	224×224	mean=[0.5], std=[0.5]	1	Preserved original thermal information without channel duplication

3.2.3 Data Augmentation Strategies

We developed a sophisticated augmentation strategy tailored specifically for thermal facial imagery, carefully balancing the need for dataset expansion with the preservation of thermally significant features:

We implemented distinct augmentation pipelines optimized for different network architectures. For RGB-designed models (AlexNet, VGG, ResNet, EfficientNet, Inception), we employed a comprehensive suite of transformations including random resized cropping, horizontal flipping, rotation (±15°), brightness and contrast adjustments (±20%), and Gaussian blurring. For our thermal-specific hybrid models, we employed a more conservative approach with grayscale conversion, random resized cropping, horizontal flipping, moderate rotation (±15°), and controlled affine transformations (±10% translation).

To address the pronounced gender imbalance in the Tufts dataset (30.32% female, 69.68% male), we implemented targeted augmentation for the underrepresented female class. This approach involved creating additional augmented samples exclusively for female subjects, effectively doubling the female representation in the training set while preserving the original male samples. This selective augmentation substantially improved class balance without introducing excessive redundancy or overfitting risks.

Our augmentation protocol was carefully calibrated to preserve the thermal signature integrity crucial for gender classification. Specifically, we avoided extreme geometric transformations and color-space alterations that might distort thermally significant facial features. The brightness and contrast adjustments were conservatively parameterized to simulate natural variations in thermal imaging conditions without introducing artifacts that could compromise the intrinsic thermal patterns.

For the combined dataset, we implemented a sophisticated integration protocol that addressed both the channel disparity between datasets and the gender distribution imbalance. To achieve perfect gender balance (50% female, 50% male), we employed controlled sampling from both constituent datasets, ensuring representative inclusion of diverse thermal imaging

conditions and subject characteristics while maintaining strict subject-level separation between training and testing partitions.

The final augmented training sets demonstrated substantially enhanced diversity and robustness. For the Tufts dataset, our class-balanced augmentation approach effectively doubled the representation of the underrepresented female class. The combined dataset benefited from both the targeted augmentation and the integration of diverse thermal imaging modalities, resulting in a comprehensive training corpus that captured a wide spectrum of thermal facial characteristics across different acquisition parameters and subject demographics.

This carefully engineered preprocessing and augmentation pipeline provided our models with high-quality, balanced training data while preserving the critical thermal signatures necessary for accurate gender classification in thermal facial imagery.

Figure 3: Thermal Image Augmentation Examples - A grid showing original thermal facial images alongside various augmented versions (horizontal flip, rotation, contrast adjustment, etc.)

Table 3: Final Experimental Dataset Configurations

Experiment	Training Set	Testing Set	Total Training Images	Total Testing Images
Tufts-only	Tufts train	Tufts test	~1,600*	~330*
Charlotte-only	Charlotte train	Charlotte test	~16,000*	~2,000*
Combined	Combined train	Combined test	18,200	2,290
Tufts-to-Charlotte	Tufts train	Charlotte test	~1,600*	~4,000*
Charlotte-to-Tufts	Charlotte train	Tufts test	~16,000*	~450*

*Approximate values after augmentation

Figure 4: Complete Data Preprocessing and Augmentation Pipeline - A flowchart showing the end-to-end process from raw dataset organization through partitioning, normalization, augmentation, to final training/testing sets.

3.3 Proposed CNN Architecture

3.3.1 Overview

Our research introduces a sophisticated deep learning framework built upon a modified ResNet-50 architecture, tailored specifically for thermal (single channel) image classification. The selection of ResNet-50 as the foundational backbone is driven by its proven ability to address challenges

inherent in training very deep neural networks. A hallmark of ResNet is its use of residual connections, which mitigate the vanishing gradient problem by introducing skip connections that allow gradients to propagate more effectively during backpropagation. This design enables the construction of deeper architectures without compromising performance, a critical advantage when extracting intricate features from complex human faces in thermal imagery.

ResNet-50 strikes an exceptional balance between computational efficiency and representational power. Its 50-layer depth facilitates the hierarchical extraction of features, ranging from low-level details such as edges and textures to high-level semantic patterns, which are essential for discerning subtle gender specific cues in thermal images. Furthermore, initializing the model with pretrained weights from ImageNet provides a robust starting point. Although thermal images differ from natural images, the general visual features learned from ImageNet—such as edge detection and texture analysis—serve as transferable knowledge that can be fine-tuned to adapt to the our domain. This transfer learning approach accelerates convergence and enhances performance, particularly when training data is limited.

The proposed architecture enhances the standard ResNet-50 by integrating three key modifications: a Channel Input Adapter to handle single-channel inputs, Squeeze-and-Excitation (SE) blocks to improve feature representation, and a redesigned classifier head to optimize classification performance. Each component is meticulously crafted to align with the implementation in the provided code, ensuring consistency between the theoretical design and practical execution.

- **Figure 5**: "Overall Architecture of HybridResNet" - A comprehensive diagram showing the complete model architecture with all components connected, highlighting the modifications to the standard ResNet-50.

3.3.2 Channel Input Adapter

Thermal imaging often presents unique challenges due to the prevalence of single-channel grayscale images, whereas pretrained models like ResNet-50 are designed for three-channel RGB inputs. To bridge this gap effectively, we developed a Channel Input Adapter that transforms single-channel inputs into a three-channel representation suitable for the pretrained backbone. Unlike the simplistic approach of replicating the grayscale channel across three dimensions, which imposes a fixed and potentially suboptimal mapping, our adapter employs a learnable transformation to capture nuanced features tailored to the input data.

Architecture

The Channel Input Adapter is implemented as a sequence of convolutional layers that progressively process the input. The transformation unfolds as follows:

- **Initial Convolutional Block**: The single-channel input, denoted as $ x \in \mathbb{R}^{1 \times H \times W} $, where $ H $ and $ W $

represent the height and width, is processed by a 3×3 convolutional layer with 32 output channels. Padding of 1 is applied to preserve spatial dimensions. This operation is followed by batch normalization to stabilize training and a ReLU activation to introduce non-linearity. The resulting feature map is $x_1 \in \mathbb{R}^{32 \times H \times W}$.

- **Subsequent Convolutional Block**: The intermediate feature map ($x1$) is fed into a second 3×3 convolutional layer, this time reducing the channel dimension to 3, again with padding of 1. Batch normalization and ReLU activation are applied subsequently, yielding the final output ($x2 \in \mathbb{R}^{3 \times H \times W}$), which matches the input requirements of the ResNet backbone.

Mathematically, the transformation can be expressed as:

$$x1 = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{3 \times 3, 32}(x)\right)\right)$$

$$x2 = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{3 \times 3, 3}(x_1)\right)\right)$$

Here, $\text{Conv}_{k \times k, c}$ represents a convolutional operation with a kernel size of $k \times k$ and c output channels, BN denotes batch normalization, and $\text{ReLU}(z) = \max(0, z)$ is the rectified linear unit activation function.

The learnable nature of this adapter allows the network to adaptively map the single-channel input to a three-channel space, potentially capturing richer and more relevant features than a static replication method. By employing convolutional layers, the adapter can learn spatially varying transformations, which is particularly advantageous for gender classification in thermal images, where local patterns—such as facial heat distributions or temperature variations—are discriminative. This design enhances the model’s compatibility with pretrained weights while optimizing its ability to process domain-specific data.

Figure 6: "Channel Input Adapter Architecture" - A detailed diagram showing the transformation from single-channel input to three-channel output, with the convolutional layers, batch normalization, and activation functions clearly labeled.

3.3.3 Squeeze and Excitation (SE) Blocks

To enhance the representational power of this model, we integrated Squeeze-and-Excitation (SE) blocks throughout the network architecture. SE blocks implement an attention mechanism that adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. This approach allows the network to increase its sensitivity to informative features while suppressing less useful ones.

The SE block operates through a two-step process: squeeze and excitation.

- **Squeeze Operation**: This step aggregates global spatial information into a channel descriptor. For convolutional feature maps (x \in \mathbb{R}^{C \times H \times W}), where (C) is the number of channels, global average pooling is applied:

[zc = \frac{1}{H \times W} \sum{i=1}^{H} \sum{j=1}^{W} xc(i, j), \quad c = 1, 2, \ldots, C]

For fully-connected layers with input (x \in \mathbb{R}^{B \times C}), where (B) is the batch size, a 1D adaptive average pooling is used:

[zc = \frac{1}{B} \sum{b=1}^{B} x_b(c)]

The result is a channel descriptor (z \in \mathbb{R}^{C}) that encapsulates the global context of each channel.

- **Excitation Operation**: The channel descriptor (z) is processed through a bottleneck structure comprising two fully-connected layers:

[s = \sigma\left(W2 \cdot \delta\left(W1 \cdot z\right)\right)]

Here, (W1 \in \mathbb{R}^{\frac{C}{r} \times C}) reduces the dimensionality with a reduction ratio (r = 16), (\delta(z) = \text{ReLU}(z)) introduces non-linearity, (W2 \in \mathbb{R}^{C \times \frac{C}{r}}) restores the original dimensionality, and (\sigma(z) = \frac{1}{1 + e^{-z}}) is the sigmoid activation function. The output (s \in \mathbb{R}^{C}) represents channel-wise attention weights ranging from 0 to 1.

- **Recalibration**: The original feature maps are scaled by these weights:
- For convolutional layers: (\tilde{x}c = sc \cdot x_c)
- For fully-connected layers: (\tilde{x}b(c) = sc \cdot x_b(c))

This recalibration enhances the emphasis on channels deemed most relevant by the attention mechanism.

Figure 7: "Squeeze and Excitation Mechanism" - Visual representation of the squeeze and excitation operations, with mathematical formulations.

We implemented SE blocks that are capable of handling both convolutional feature maps (4D tensors) and fully-connected layers (2D tensors), making the architecture more flexible. For convolutional layers, the SE blocks apply 2D adaptive average pooling before computing attention weights, while for fully-connected layers, they utilize 1D pooling. This adaptive approach ensures that the attention mechanism works effectively throughout the network.

SE blocks are integrated into the ResNet architecture by appending them after the final convolutional layer (conv3) of each bottleneck module within layers 1 through 4. This strategic placement ensures that feature

recalibration occurs at multiple abstraction levels, enhancing the model's ability to prioritize features critical for the classification task.

The standard ResNet classifier, consisting of a single fully-connected layer, is replaced with a more elaborate structure to optimize classification performance and generalization. This redesign addresses the need for robust feature processing and regularization, particularly in the context of gender classification in thermal images.

The classifier head processes the 2048-dimensional feature vector obtained from global average pooling through a multi-layer sequence:

- **First Dropout Layer**: A dropout operation with a probability of 0.5 is applied to the input (x \in \mathbb{R}^{2048}), randomly setting half of the features to zero during training to prevent neuron co-adaptation.
- **Dimensionality Reduction**: A fully-connected layer reduces the dimensionality from 2048 to 512, followed by a ReLU activation:

[x2 = \text{ReLU}(W1 x1 + b1)]

where (W1 \in \mathbb{R}^{512 \times 2048}) and (b1 \in \mathbb{R}^{512}) are learnable parameters.

- **SE Block**: An SE block recalibrates the 512-dimensional feature vector, applying the squeeze and excitation operations described earlier to emphasize discriminative features.
- **Second Dropout Layer**: Another dropout operation with a probability of 0.3 provides additional regularization.
- **Output Layer**: A final fully-connected layer maps the features to the number of classes:

[y = W2 x4 + b_2]

where (W2 \in \mathbb{R}^{\text{num_classes} \times 512}) and (b2 \in \mathbb{R}^{\text{num_classes}}) produce the classification logits.

The full transformation is:

[x1 = \text{Dropout}{0.5}(x)]

[x2 = \text{ReLU}(W1 x1 + b1)]

[x3 = \text{SEBlock}(x2)]

[x4 = \text{Dropout}{0.3}(x_3)]

[y = W2 x4 + b_2]

The incorporation of SE blocks enhances the network's sensitivity to informative features, a crucial capability in gender classification for thermal

images, where subtle differences in facial heat distribution can be discriminative. The adaptive nature of the attention mechanism allows the model to dynamically adjust its focus, improving both performance and robustness across diverse datasets.

Figure 9 - Visual comparison between standard ResNet and your HybridResNet highlighting the key differences.

3.4 Model Comparison

This section provides an evaluation of the diverse neural network architectures employed as baseline models to compare with our proposed ResNet-based framework for gender detection in thermal facial images. The selection criteria prioritized architectural diversity across model generations, parameter complexity, and feature extraction methodologies to establish a robust comparative foundation. By examining AlexNet, VGG-16, InceptionV3, ResNet50, and EfficientNet, we gain valuable insights into how different architectural paradigms process the unique characteristics of thermal facial signatures for gender classification tasks.

3.4.1 Baseline Architectures

3.4.1.1 AlexNet: Foundational CNN Architecture

AlexNet represents a fundamental benchmark in our evaluation due to its historical significance in revolutionizing computer vision through deep convolutional neural networks. Despite its relative simplicity by contemporary standards, this architecture offers critical insights into the minimum viable model complexity required for effective thermal feature discrimination. The network comprises five convolutional layers and three fully connected layers, creating a relatively shallow architecture with eight trainable layers.

The model processes thermal input images resized to 224×224 pixels through a series of operations beginning with large-kernel convolutions (11×11 stride 4) that capture broad thermal gradients across facial regions. These initial layers are particularly relevant for thermal imaging, as they can detect coarse temperature variations corresponding to major facial vasculature patterns that exhibit gender-specific differences. Subsequent layers employ progressively smaller kernels (5×5, then 3×3) to refine feature representation, with max-pooling operations providing spatial reduction. The final network outputs a 4096-dimensional feature vector before classification, which must encapsulate gender-discriminative thermal signatures.

AlexNet's inclusion allows us to evaluate whether early CNN architectural patterns can correctly capture the subtle temperature distribution differences between male and female thermal facial signatures. The model's Local Response Normalization (LRN) layers may also prove beneficial in standardizing thermal intensity variations across different capture conditions.

3.4.1.2 VGG-16: Homogeneous Deep Architecture

VGG-16 extends architectural depth systematically through homogeneous convolutional blocks, enabling examination of how increased layer count (16 trainable layers) affects thermal feature learning without introducing advanced structural innovations. Its uniform design philosophy—consisting of stacked 3×3 convolutional layers followed by spatial reduction via max-pooling—provides a controlled comparison point for evaluating thermal feature learning in deeper networks.

The network's consistent kernel size (3×3) throughout all convolutional layers creates a large effective receptive field while maintaining computational efficiency. This architecture may effectively capture multi-scale thermal patterns ranging from localized temperature peaks around the periorbital regions to broader thermal distributions across facial contours.

VGG-16's straightforward layer progression offers interpretability advantages, potentially allowing clearer attribution of which facial thermal regions contribute most significantly to gender classification decisions. This transparency could prove valuable for subsequent research into the physiological basis of gender-specific thermal signatures.

3.4.1.3 InceptionV3: Multi-Scale Processing Architecture

InceptionV3 introduces sophisticated multi-scale processing capabilities through its innovative inception modules and factorized convolutions. This architectural approach allows simultaneous analysis of thermal features at multiple spatial resolutions—a potentially valuable characteristic for thermal gender classification, where discriminative information may exist at different scales, from fine vascular patterns to broader facial temperature zones.

The architecture reduces computational complexity through strategic use of 1×1 convolutions for dimensionality reduction prior to expensive 3×3 and 5×5 operations. Its asymmetric kernel decompositions (replacing 5×5 filters with stacked 3×3 convolutions and factorizing n×n filters into consecutive 1×n and n×1 operations) enhance efficiency while preserving representational capacity. Input thermal images are resized to 299×299 pixels to align with InceptionV3's native resolution, providing increased spatial detail compared to other models in our evaluation.

InceptionV3's auxiliary classifier, which emerges from an intermediate layer during training, potentially aids in propagating more direct gender-classification signals to earlier network stages. This feature may prove particularly beneficial for thermal imaging, where distinguishing gradient information can be more subtle than in visible-spectrum imagery.

The network's branch diversity within inception modules enables it to learn specialized feature extractors for different thermal pattern types simultaneously—potentially capturing both the textural aspects of facial

thermal patterns and their spatial configuration in a single unified architecture.

3.4.1.4 ResNet50: Residual Learning Framework

ResNet50 employs innovative residual learning principles to achieve 50-layer depth without degradation in training accuracy. The architecture utilizes bottleneck blocks to mitigate vanishing gradients, enabling deeper feature hierarchies that may better capture the complex relationships in thermal facial imagery.

Each residual block follows the fundamental mapping principle:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, {W_i}) + \mathbf{x}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent input and output vectors respectively, and $\mathcal{F}$ denotes the residual function implemented through three consecutive convolutions (1×1 for dimensionality reduction, 3×3 for feature extraction, and 1×1 for restoration). This design enables stable training of substantially deeper networks while preserving gradient flow—a critical consideration when fine-tuning on limited thermal datasets where gradient signals may be weaker than in large-scale visible image collections.

The architecture's identity shortcuts create direct paths for gradient flow, potentially enabling more effective training on the relatively subtle thermal features that distinguish genders. The bottleneck design (1×1, 3×3, 1×1 convolution pattern) reduces computational requirements while maintaining representational capacity, making ResNet50 an efficient option for thermal image processing despite its depth.

ResNet50's batch normalization after each convolutional layer standardizes feature activations, which may be particularly beneficial for thermal imagery where absolute temperature values can vary across subjects and capture conditions. This normalization potentially helps the network focus on relative temperature distribution patterns rather than absolute values.

3.4.1.5 EfficientNet-B0: Optimized Scaling Architecture

EfficientNet-B0 represents the state-of-the-art in efficiency-optimized architectures, leveraging compound scaling to balance depth, width, and resolution. This balanced approach optimizes the accuracy-efficiency trade-off, making it particularly relevant for potential deployment of thermal gender classification systems in resource-constrained environments.

The architecture employs mobile inverted bottleneck (MBConv) layers as its primary building block, integrating squeeze-and-excitation (SE) mechanisms for adaptive channel attention. This attention mechanism can be formulated as:

$$\mathbf{\tilde{X}} = s(\mathbf{X}) \cdot \mathbf{X}$$

where $s(\mathbf{X})$ represents channel-wise attention weights derived from the SE block. This mechanism potentially enables the network to emphasize the most gender-discriminative thermal channels while suppressing less informative ones, creating an adaptive feature selection process beneficial for capturing subtle thermal differences between genders.

Despite its relatively lightweight design (5.3 million parameters), EfficientNet-B0 achieves competitive accuracy on standard computer vision benchmarks. This raises the important question of whether efficient architectures can maintain high accuracy on thermal gender classification despite their reduced capacity, or if the subtle nature of thermal features requires larger models. The architecture's depth (82 layers) combined with its parameter efficiency provides an interesting contrast to other models in our evaluation.

The swish activation function ($x \cdot \text{sigmoid}(x)$) used throughout EfficientNet potentially offers advantages over ReLU activations for thermal imagery by providing smoother gradients for small activation values, which may better preserve subtle thermal variation information during forward propagation.

Comparison of Architectural Characteristics

Table 4: Comprehensive Baseline Model Specifications

Model	Depth	Parameters (M)	Input Size	Key Components	Potential Thermal Imaging Advantages
AlexNet	8	61.0	224×224	Large kernels (11×11), LRN	Effective capture of broad thermal gradients
VGG-16	16	138.0	224×224	Homogeneous 3×3 conv stacks	Consistent multi-scale feature extraction
InceptionV3	48	23.9	299×299	Factorized convolutions, Auxiliary classifier	Multi-resolution thermal pattern analysis
ResNet50	50	25.6	224×224	Bottleneck residual blocks	Deep thermal feature hierarchies with gradient preservation
EfficientNet-B0	82	5.3	224×224	MBConv with SE, Compound scaling	Adaptive attention to gender-discriminative

Model	Depth	Parameters (M)	Input Size	Key Components	Potential Thermal Imaging Advantages
					thermal channels

Figure 10: Architectural Diagrams – Detailed schematic representations of each baseline model's layer configuration, highlighting specific components relevant to thermal feature extraction.

3.4.2 Input Adaptation and Training Protocol

3.4.2.1 Thermal Image Preprocessing and Channel Adaptation

Adapting standard CNN architectures designed for RGB images to thermal data requires careful consideration of input channel dimensionality. Our experimental protocol addresses this challenge through dataset-specific preprocessing pipelines:

For the Charlotte dataset's single-channel thermal inputs, we employed channel replication to create compatible three-channel inputs. This approach converts grayscale thermal intensity values into three identical channels during image loading, preserving the original thermal distribution while satisfying the input requirements of networks pretrained on RGB data. While this method introduces redundancy, it maintains compatibility with the convolutional filters learned from visible spectrum imagery and allows us to leverage transfer learning effectively.

The Tufts dataset provides native three-channel thermal representations, each encoding different thermal wavelength bands. These original multi-channel thermal representations were retained and directly utilized without modification to preserve the potential complementary information across thermal spectral bands. The three-channel structure of this data aligns naturally with the input expectations of conventional CNNs.

3.4.2.2 Transfer Learning and Fine-Tuning Strategy

We initialize the baseline models with ImageNet-pretrained weights, freezing their initial layers to retain general feature extraction while training only a new final classification layer for thermal gender classification. This focuses optimization on our specific two-class task and minimizes overfitting.

3.4.3 Benchmarking Objectives

Our comparative analysis framework employs a multi-faceted evaluation approach. The benchmarking objectives address two critical dimensions:

1. Classification Performance Metrics

Beyond standard accuracy, we evaluate models using metrics particularly relevant to biometric applications:

- Confusion matrices: To identify gender-specific misclassification patterns.
- F1-score: Calculated per class via classification reports to balance precision and recall considerations.

These metrics are calculated across both individual datasets (Charlotte and Tufts separately) and combined data scenarios to assess model robustness under varying thermal imaging conditions and demographic distributions.

2. Cross-Dataset Generalization Assessment

A critical challenge in thermal imaging for gender detection is generalization across different sensor technologies, environmental conditions, and demographic compositions. We conduct rigorous cross-dataset evaluations:

- **Tufts-to-Charlotte:** Training on the Tufts dataset (with its three-channel representation) and evaluating on the Charlotte dataset.
- **Charlotte-to-Tufts:** Training on the Charlotte dataset (with single-channel thermal data) and evaluating on the Tufts dataset.

3.5 Experimental Setup

To rigorously assess the efficacy of our baseline models and the proposed hybrid architecture, we designed a comprehensive experimental framework involving multiple dataset configurations. Specifically, we utilized the Tufts dataset, the Charlotte dataset, a combined dataset merging both, and two cross-dataset scenarios: training on Tufts and testing on Charlotte (Tufts-to-Charlotte), and vice versa (Charlotte-to-Tufts). These configurations enabled us to evaluate the models' performance within individual datasets as well as their ability to generalize across distinct datasets, a critical aspect of real-world applicability.

For training, all models were optimized using the Adam algorithm, configured with momentum parameters (\beta1 = 0.9) and (\beta2 = 0.999), which are widely adopted for their stability and efficiency in deep learning tasks. We set the initial learning rate to 0.00005, a value selected to ensure gradual parameter updates suitable for our architecture. To enhance training dynamics, we implemented a 5-epoch warmup phase during which the learning rate increased linearly from zero to the specified value, followed by cosine annealing for the subsequent epochs to promote smooth convergence to an optimal solution. We experimented with batch sizes of 32 and 64 to explore their effects on training stability and generalization performance, providing insights into the trade-offs between computational efficiency and model accuracy.

Each model underwent training for 10 epochs, a duration determined through preliminary experiments to strike a balance between achieving convergence and minimizing computational overhead. The experiments were executed on an NVIDIA GeForce RTX 4090, a high-performance hardware platform that facilitated rapid iteration. To optimize data handling and reduce training bottlenecks, we employed PyTorch's DataLoader with settings of `num_workers=8` and `pin_memory=True`, ensuring efficient data transfer to the GPU and maximizing throughput during training.

4. Experimental Results

- **Length**: 5-6 pages.
- **Content**:
  - **4.1 Experimental Setup**:
    - Hardware: GPU specifications (e.g., NVIDIA RTX 3090).
    - Software: Frameworks (e.g., PyTorch, TensorFlow), libraries.
    - Evaluation metrics: Accuracy, precision, recall, F1-score.
    - **Explanation**: Describe train-test split or cross-validation strategy.
  - **4.2 Results on Individual Datasets**:
    - **4.2.1 Tufts Dataset**:
      - **Table**: "Table 5: Performance on Tufts Dataset" (columns: Model, Accuracy, Precision, Recall, F1).
      - **Explanation**: Analyze top performers and why (e.g., deeper models handle limited features better).
    - **4.2.2 Charlotte Dataset**:
      - **Table**: "Table 6: Performance on Charlotte Dataset" (columns as above).
      - **Explanation**: Compare with Tufts, note differences due to channel availability.
  - **4.3 Results on Combined Dataset**:
    - **Table**: "Table 7: Performance on Combined Dataset" (columns as above).
    - **Explanation**: Discuss improvements or challenges from combining datasets.
  - **4.4 Proposed Model Performance**:
    - **Table**: "Table 8: Proposed Model vs. Baselines" (columns: Dataset, Model, Accuracy, F1).
    - **Explanation**: Highlight advantages (e.g., channel adapter, SE blocks).
  - **4.5 Ablation Study**:
    - Components tested: Channel adapter, SE blocks.
    - **Table**: "Table 9: Ablation Study Results" (columns: Configuration, Accuracy, F1).
    - **Explanation**: Justify inclusion of each component based on performance drop without them.
  - **4.6 Visualizations**:
    - **Diagram**: "Figure 4: Confusion Matrices" (show for proposed model on each dataset).

        - **Diagram**: "Figure 5: ROC Curves" (compare proposed model vs. best baseline).
        - **Explanation**: Discuss correct/incorrect prediction examples with sample images.
- **Instructions**:
    - Present tables immediately after their subsections for easy reference.
    - Use Figures 4-5 to visually support the quantitative results.
    - Provide detailed analysis after each table/diagram (1-2 paragraphs).

---

5. Discussion

- **Length**: 3-4 pages.
- **Content**:
    - **5.1 Implications of Findings**:
        - Advancement in thermal gender classification.
        - **Explanation**: Discuss real-world applications (e.g., security, automotive).
    - **5.2 Challenges and Limitations**:
        - Thermal imaging limitations: Lack of facial detail.
        - Dataset issues: Class imbalance, channel differences.
        - Computational constraints: Training time, resource demands.
        - **Explanation**: Provide specific examples from results (e.g., lower recall for females).
    - **5.3 Future Directions**:
        - Explore multimodal approaches (thermal + visible).
        - Enhance architecture with transformers or other techniques.
        - **Explanation**: Suggest how these could address current limitations.
- **Instructions**:
    - Use narrative style to connect results to broader context.
    - Avoid introducing new data; focus on interpreting Section 4.

---

6. Conclusion

- **Length**: 1-1.5 pages.
- **Content**:
    - Summarize key findings: Performance of baseline models, success of proposed architecture.
    - Reiterate contributions: Comprehensive evaluation, novel CNN design.
    - Emphasize significance: Robust gender detection in challenging conditions.
    - Suggest next steps: Larger datasets, real-time implementation.
- **Instructions**:
    - Keep concise but impactful, reinforcing the paper’s value.

References

- **Length**: 1-2 pages.
- **Content**:
  - List all cited works (aim for 30-40 references).
  - Include studies from Sections 2-4, dataset papers, and deep learning references.
- **Instructions**:
  - Use a consistent citation style (e.g., APA, IEEE).

Appendices (Optional)

- **Length**: 1-2 pages (if included).
- **Content**:
  - Additional preprocessing details.
  - Full hyperparameter tables.
  - Code snippets (e.g., proposed model implementation).
- **Instructions**:
  - Include only if space allows and content enhances understanding.

Final Notes

- **Total Length**: This outline targets 20 pages by allocating:
  - Abstract: 0.5 page
  - Introduction: 2 pages
  - Related Work: 4 pages
  - Datasets and Methodology: 6 pages
  - Experimental Results: 6 pages
  - Discussion: 4 pages
  - Conclusion: 1.5 pages
  - References: 2 pages (adjust as needed).
- **Visual Elements**: Include at least 5 figures (diagrams) and 9 tables, placed strategically to break up text and enhance readability.
- **Writing Tips**:
  - Expand explanations with examples, equations (e.g., SE block math), and detailed analyses.
  - Use subheadings to maintain structure and guide the reader.
  - Ensure each section flows logically into the next, referencing earlier sections where relevant.

This outline provides a robust framework to write an extensive, informative research paper. Follow the instructions for each section to ensure depth and clarity, and adjust content as needed during writing to meet the 20-page goal.