# Towards All-in-One Medical Image Re-Identification

Harathi Boddu
MS in Computer and Information Sciences,
Texas Tech University,
Lubbock, Texas, USA.

Rajasri Kondaveeti
MS in Computer and Information Sciences,
Texas Tech University,
Lubbock, Texas, USA.

Hari Krishna Cherukuri
MS in Computer and Information Sciences,
Texas Tech University,
Lubbock, Texas, USA.

Sai Sanjay Chitteni
MS in Computer and Information Sciences,
Texas Tech University,
Lubbock, Texas, USA.

Rupesh Kalyanam
MS in Computer and Information Sciences,
Texas Tech University,
Lubbock, Texas, USA.

*Abstract*— **Accurate patient identification from images is important to correct diagnosis and healthcare delivery. This task is more challenging as images are captured across diverse modalities such as CT, MRI, X-ray, and microscopic images. Existing person ReID models generally employ a single global feature extractor, which hinders how much modality-specific features they can extract, causing accuracy and flexibility to suffer. We present here a multimodal deep learning system that responds to these problems by adding modality-aware classification and modality-specific convolutional neural networks. The process first classifies the modality of the input image and then forwards it to a particular CNN that is specialized for the modality. The process also uses advanced preprocessing methods to enrich structural features and real-time data augmentation to improve generalization. This module-based architecture strives to provide a more accurate and efficient solution for patient identification on various imaging modalities.**

*Keywords*—**Medical Image Re-identification, Personalized Healthcare, Modality Adaptation, Foundation Models, medical imaging, patient re-identification, CNN, multimodal deep learning, COMPA, EfficientNet**

## I. Introduction

Medical imaging plays a crucial role to clinical diagnosis, treatment planning, and follow-up of patients in the long term. With the use of imaging modalities like MRI, CT, X-ray, and microscopic images comprising nearly 90% of all medical data, the ability to be able to correctly identify and track patients through multiple scans is becoming more critical than ever. The existing medical systems, however, are founded on metadata—file name or patient ID—to retrieve prior images. These identifiers are incomplete or lost, particularly when data are shared across different storage systems or organizations. This presents a twofold challenge: the inability to correctly re-identify patients across modalities and growing concerns regarding privacy breaches due to the inclusion of subtle visual cues within medical images.

Existing medical image re-identification (MedReID) approaches typically rely on unified models trained across multiple modalities. However, such methods often suffer from reduced accuracy, as a single model struggles to generalize across the diverse visual and structural characteristics inherent to different imaging types. While recent models like MaMI have explored modality-adaptive learning, they still face limitations when scaling across new modalities or dealing with high intra-class variability in patient images.

To address these gaps, we propose a novel multimodal deep learning framework designed specifically for patient re-identification across medical images. Our approach employs a modular architecture beginning with a modality classification model (COMPA), which accurately detects the imaging modality using architectures like ResNet50 or EfficientNet. Based on this prediction, each image is routed to a dedicated convolutional neural network optimized for that modality—ensuring more accurate and relevant feature extraction. To further enhance performance, we incorporate advanced preprocessing techniques, including contrast enhancement, noise reduction, and edge detection, along with real-time data augmentation to improve model generalization.

Our method is evaluated on a diverse set of publicly available datasets representing different modalities and body regions. Experimental results show that our framework significantly improves both identification accuracy and computational efficiency while maintaining strong privacy-preserving characteristics. This positions our system as a practical and scalable solution for real-world MedReID applications in clinical environments.

New approaches like MaMI have evolved by suggesting modality-adaptive re-identification architectures in the medical field. They typically operate on advanced transformer-based models and continuous parameter fine-tuning, which render them computationally expensive and less interpretable. More importantly, they don't result in explicit modular separation between patient identification and modality detection, such that they suffer from performance collapse when applied to novel or heterogeneous datasets. Our framework fills these gaps by presenting a completely modular, end-to-end deep learning pipeline tailored to medical image-based patient re-identification. The central idea behind our design is to separate the identification task of imaging modality from the identification task of patient, enabling each phase to be specialized and optimized in isolation.

The first module in our pipeline is COMPA—a lightweight modality classification model trained using transfer learning architectures such as ResNet50 and EfficientNet-B0. COMPA classifies whether the input image is of CT, MRI, X-ray, or microscopic modality. Once the modality is detected, the image is passed through a corresponding CNN model for that specific modality so that feature extraction is enhanced without cross-modality noise.

In order to supplement the quality of the input image, we use a modality-aware preprocessing chain including grayscale conversion, CLAHE, Gaussian noise removal, and Canny edge detection. These processes are used to emphasize structural and anatomical features relevant to patient identification. Real-time data augmentation in rotation, flipping, and intensity modifications is also included in our system for increased training variability and overfitting avoidance, very much an important factor given the sparsity of the size of labeled medical data. All CNNs within our system are supervised, trained with classification and patient-wise labeling, where each output class corresponds to a unique patient. It is additionally optimized by categorical cross-entropy loss. Modular design allows CNNs to be trained, tested, and updated independently, allowing for flexibility and scalability when new modalities or patient data become available.

Unlike other integrated systems, our system can be compatible with real-time clinical deployment through a Flask-based web interface that supports efficient usability in healthcare environments. The system is lean, efficient, and is capable of running on mid-range hardware without taking the expense of high-end GPU infrastructure. Briefly, our model presents a realistic, interpretable, and high-quality medical image re-identification solution. By learning processing pipelines for particular modalities and emphasizing robust preprocessing and augmentation, we greatly improve identification quality without sacrificing computational efficiency. This renders our approach a promising candidate for real-world patient tracking, individualized treatment, and secure medical data management in clinical processes.

## II. RELATED WORK

Identity Retrieval in Medical Imaging: Unlike disease detection or segmentation, the task of retrieving patient identity from medical images remains a relatively underexplored area. Most conventional approaches rely on explicit metadata (e.g., patient names or study IDs) rather than visual features. A few early attempts utilized shallow image descriptors or edge-based features on CT or X-ray images for identity matching, but these methods failed to scale across modalities or account for visual variability between scans. With the growing availability of large-scale multimodal datasets, the potential for visually driven patient identification is becoming more feasible—yet remains underutilized.

Gaps in Visual Re-Identification for Clinical Use: Visual re-identification has been extensively researched in domains such as person tracking, facial recognition, and vehicle monitoring, where appearance varies due to pose, lighting, or occlusion. However, these models assume high inter-class variability and context-rich features, unlike medical images where variations are subtle and often diagnostic in nature. Furthermore, unlike in surveillance applications, medical re-identification must avoid altering or masking clinically relevant features, creating a unique set of design constraints.

Emergence of Generalized Medical Models: Recent progress in vision-language and foundation models has influenced the medical imaging field, leading to models pretrained on large collections of labeled or weakly labeled clinical data. These foundation models have shown promise in diagnosis, report generation, and image classification across X-ray, CT, and retinal images. However, their application to re-identification tasks has been minimal, as they are typically optimized for pathology detection rather than subtle inter-patient differences. Our work bridges this gap by aligning pretrained representations with identity-preserving objectives.

Need for Modality-Aware Frameworks: Most medical deep learning pipelines treat all input images uniformly, regardless of imaging modality. This limits their capacity to extract relevant structural or textural features unique to CT, MRI, or microscopic images. While some works have experimented with modality-specific finetuning, a systematic mechanism for routing and processing images based on modality remains rare. A modular, adaptive approach—

where each modality is handled by a specialized network—can significantly improve accuracy and flexibility, especially in heterogeneous clinical datasets.

While modality-agnostic models like MaMI have introduced mechanisms to account for imaging variations with transformer-based backbones, they are computationally costly and non-interpretable—significant roadblocks to clinical adoption in real-world settings. Unified models are vulnerable to cross-modal feature entanglement, which tends to impede correct identity discrimination, particularly in cases of weak and non-overlapping modality-specific anatomical patterns. Moreover, being monolithic, their scalability or extension to new modalities incurs significant retraining.

Recent advances in medical AI have seen the rise of foundation models and self-supervised learning techniques, including Vision Transformers and models like BioMedCLIP. While effective for tasks such as diagnosis, segmentation, and report generation, these models are not optimized for patient identity retrieval. Most rely on learning disease-related patterns rather than subtle biometric cues. Our approach addresses this gap through a modular architecture that explicitly separates modality classification, tailored preprocessing, and modality-specific CNN-based identification. This design improves both accuracy and transparency, while supporting flexible integration into clinical workflows where adaptability, interpretability, and deployment readiness are essential.

## III. APPROACH.

### A. Overview

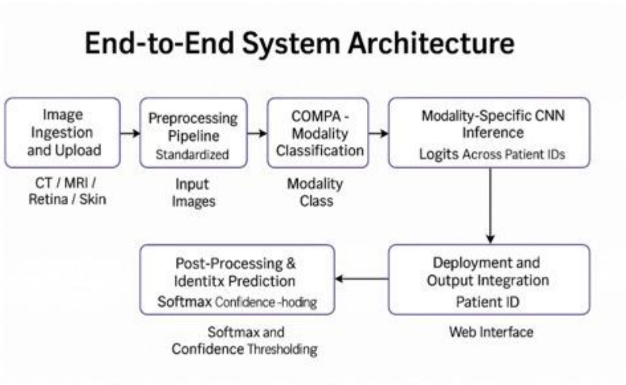
End-to-End System Architecture
Image Ingestion and Upload
CT / MRI / Retina / Skin
Preprocessing Pipeline
Standardized
Input Images
COMPA - Modality Classification
Modality Class
Modality-Specific CNN Inference
Logits Across Patient IDs
Deployment and Output Integration
Patient ID
Web Interface
Post-Processing & Identitx Prediction
Softmax Confidence -hoding
Softmax and Confidence Thresholding

Figure 1. Overview of the Proposed End-to-End System Architecture

The proposed system follows a modular, end-to-end architecture designed for multimodal patient re-identification from medical images. The entire pipeline consists of six primary stages: image ingestion, preprocessing, modality classification, CNN inference, post-processing, and deployment. Each component is optimized to handle diverse input modalities, including CT, MRI, retinal (fundus), and dermoscopic skin images.

The process begins with the Image Ingestion and Upload module, where raw medical images of various modalities are uploaded through the web interface. These inputs are first directed to a Preprocessing Pipeline, which applies standardized operations such as grayscale conversion, CLAHE for local contrast enhancement, Gaussian noise filtering, and Canny edge detection. This preprocessing step ensures that all inputs are uniform and noise-reduced, providing better feature clarity for downstream tasks.

The preprocessed image is then passed to the COMPA Modality Classification module, a lightweight deep learning model trained to identify the image's modality (e.g., CT, MRI, retina, skin). COMPA outputs a modality class, which determines the routing path for further processing. Once the modality is identified, the image is routed to the corresponding Modality-Specific CNN Inference model. Each CNN is independently trained on a specific imaging modality and is responsible for predicting the patient identity. The model outputs a logit vector representing class probabilities across all known patient IDs for that modality.

Following inference, Post-Processing and Identity Prediction is carried out by applying a softmax activation function and confidence thresholding to finalize the prediction. This ensures that only high-confidence predictions are passed forward to the interface layer. The final stage, Deployment and Output Integration, delivers the predicted patient ID to the Web Interface, where it is displayed to the user. This modular deployment strategy ensures that the system remains lightweight, interpretable, and easily upgradable as more modalities or patient classes are added in the future. This architecture enables real-time, accurate patient re-identification while remaining robust to modality variations, dataset imbalances, and clinical deployment constraints.

Our proposed system addresses the complex task of patient re-identification across diverse medical imaging modalities through a modular, multimodal deep learning framework. Unlike conventional approaches that employ a single unified model, our system is divided into distinct stages: modality classification, preprocessing, modality-specific CNN-based identification, and web-based deployment for real-time inference. This architecture allows the system to specialize its operations based on the modality of the input image, thereby improving identification accuracy and system flexibility.

The pipeline begins with a modality classification module, COMPA, utilizing transfer learning for detection of whether the input scan belongs to CT, MRI, X-ray, or microscopic modality. Based on this prediction, the image is sent to a modality-specific CNN that has been trained for that particular modality. Sophisticated preprocessing techniques such as CLAHE, Gaussian noise removal, and edge detection are utilized to enhance image features before classification. Apart from that, data augmentation techniques are used to make the model more robust and generalized. The final patient ID prediction is delivered through a lightweight Flask-based web interface designed for deployment in clinical environments.

### *B. Modality Detection using COMPA*

A key innovation of our framework is the use of the COMPA (Classifier Of Medical image PAtterns) model for modality classification. This model serves as the entry point of the system and determines the appropriate processing path for each image. COMPA is built on state-of-the-art transfer learning architectures i.e. ResNet50 and EfficientNet-B0, that are transferred to a labeled image dataset of four modalities. The final layer has four output classes corresponding to CT, MRI, X-ray, and microscopic images and incorporates a softmax classifier.

This approach ensures rapid and accurate modality classification without the need for complex parameter-adaptive methods, such as is the case with transformer-based models. Second, by decoupling modality detection from patient detection, COMPA allows the system to remain interpretable and easily scalable in the future when integrating new modalities.

Preprocessing and Image Enhancement: Because of the varying visual qualities and properties of medical images across modalities, preprocessing is required to bring out the important identity-related features. The images are first converted into grayscale to focus on structural content as well as reduce dimensionality. CLAHE (Contrast Limited Adaptive Histogram Equalization) is then used to locally enhance contrast, particularly in low-intensity regions. To reduce noise in imaging, a Gaussian filter is applied to blur the image while preserving important edge information.

Canny edge detection is subsequently employed to achieve high-gradient boundaries that are likely to portray the most significant anatomical structures. The edges are further pseudo-colored with overlay to highlight significant contours. Preprocessing pipeline strives to normalize and enhance the input data such that it is simpler for the CNNs to learn discriminative and informative features for patient identification.

Once modality detection is done, the image is then fed to a dedicated CNN model specifically trained to detect patients in the detected modality. This modular solution contrasts with existing unified models that use all modalities on a single network, leading to suboptimal performance due to modality interference. For CT and X-ray images, we use EfficientNet-B0, a lightweight yet powerful model that balances speed and accuracy. MRI scans, which contain more complex patterns and finer gradients, are handled by a deeper ResNet50 architecture. Microscopic images, such as those used in histopathology or retinal scans, are processed using a hybrid model that combines the strengths of ResNet and EfficientNet. Each of these CNNs is trained using supervised learning with patient-wise labels, allowing the final softmax layer to output a unique patient ID prediction for each input.

To mitigate data sparsity and class imbalance—usual challenges in medical imaging—we used a systematic data augmentation technique. At training time, the input images undergo a sequence of real-time operations such as random rotation, flips in the horizontal and vertical directions, zoom, shift, brightness, and contrast adjustments.

These augmentations imitate the natural variability of actual clinical data, e.g., imaging angles or lighting conditions. This not only prevents overfitting but also increases the model's ability to generalize across patients and image devices. Data augmentation adds controlled randomness and raises the model's resilience without requiring additional labeled samples.

Once an image has been processed by its respective CNN, the model outputs a predicted patient ID using a softmax classifier. The CNNs are trained as multi-class classifiers, with each class representing a unique patient. In cases where retrieval-based identity matching is preferred, embedding-based distance metrics such as cosine similarity can be optionally applied. This component can be adapted based on whether the deployment scenario requires classification or matching.

The integration of modality detection, preprocessing, and identity classification is tightly coordinated, ensuring seamless operation from input to output. The modular nature of the pipeline allows for easy replacement or retraining of individual components without disrupting the entire system. To support real-world clinical deployment, the entire framework is encapsulated in a Flask-based web application. The frontend, built using HTML, CSS, and JavaScript, allows users to upload medical scans. The backend handles image preprocessing, modality detection using COMPA, routing to the correct CNN model, and prediction display. This architecture supports local and cloud deployment, making it suitable for both research and operational use. Its lightweight design ensures compatibility with moderate computational resources, eliminating the need for expensive GPU infrastructure in production settings. Our approach offers significant advantages over existing systems. The use of supervised modality classification simplifies routing and increases interpretability. The separation of modality-specific models allows for better specialization and scalability.

Advanced preprocessing techniques enhance image quality and focus model attention on key features, while extensive data augmentation strengthens generalization. Finally, the integration into a lightweight web application ensures usability in real-time clinical environments.

### *C. Loss Functions and Training Objectives:*

Each CNN model in our framework is trained using a supervised learning approach, with the objective of minimizing the categorical cross-entropy loss. This loss function is widely used for multi-class classification tasks such as patient identity prediction.

The **categorical cross-entropy loss** is defined as:

$$L_{CE}= \sum_{i=1}^{n} ylog(y1)$$

Where:

y is the ground-truth one-hot encoded label for class

y1 is the predicted probability for class,

n is the total number of classes (patients in this case).

This loss encourages the model to output high confidence for the correct class while penalizing incorrect predictions. Each modality-specific CNN uses this loss independently, enabling tailored optimization for identity classification.

### *D. Computational Complexity and Deployment Suitability:*

Compared to existing transformer-based systems like MaMI, our modular CNN-based architecture is computationally lighter. The COMPA classifier and modality-specific CNNs all use architectures that can be efficiently deployed on CPUs or mid-tier GPUs. The model sizes range between 5–25MB per CNN, and inference times are under 100ms per image on a standard machine. This design ensures feasibility for deployment in real-time hospital systems and integration into PACS (Picture Archiving and Communication Systems).

## IV. EXPERIMENTS

### *A. Model Details*

The architecture of our proposed system is composed of two main deep learning modules: the COMPA model for modality classification and a set of modality-specific CNN models for patient identity recognition. This modular approach ensures that each model is tailored to the specific characteristics of its input modality, thereby improving accuracy and enabling scalable deployment in clinical environments.

The COMPA (Classifier Of Medical image PAtterns) module serves as the initial stage in the processing pipeline. Its role is to accurately classify the imaging modality of the input scan, such as CT, MRI, X-ray, or microscopic images. COMPA is implemented using transfer learning by fine-tuning well-established deep learning architectures like ResNet50 and EfficientNet-B0, both pretrained on the ImageNet dataset. The input to the model is a resized image of dimensions 224×224×3224 \times 224 \times 3224×224×3, and the final layers are customized to suit the classification task. After the convolutional backbone, a Global Average Pooling layer is applied, followed by a dense layer with 256 ReLU-activated units and a dropout layer with a dropout rate of 0.3. The output layer is a softmax classifier with four units corresponding to the four modality classes. The model is trained using categorical crossentropy loss, optimized via the Adam optimizer with a learning rate of 0.0001, and evaluated using accuracy as the primary performance metric. Early stopping and model checkpointing are applied to prevent overfitting and to retain the best-performing model.

Once the modality is predicted, the input image is forwarded to one of four specialized CNN models, each trained to recognize patient identity within a specific modality. For CT and X-ray images, we use EfficientNet-B0 due to its lightweight architecture and competitive performance on high-contrast grayscale images. These models are initialized with pretrained ImageNet weights and fine-tuned on patient-labeled CT and X-ray datasets. The architecture includes a Global Average Pooling layer, followed by a dense layer with 512 ReLU units, a dropout layer with a rate of 0.4, and a softmax output layer whose size matches the number of unique patient identities in the dataset. MRI images, which are typically more complex due to the presence of soft tissue textures and subtle anatomical features, are processed using a deeper ResNet50 model. The same architectural modifications as used in EfficientNet-B0 are applied here, including global pooling, fully connected layers, and dropout for regularization. The model is capable of learning fine-grained features essential for distinguishing between patients, even in cases with minimal inter-class variation. In contrast, microscopic images, such as histopathology or retinal scans, require the model to capture both global context and fine local textures. To handle this, we designed a hybrid model that combines the feature extraction capabilities of ResNet50 and EfficientNet-B0. The intermediate features from both networks are concatenated and passed through a fully connected layer, followed by dropout and a final softmax layer. This hybrid architecture performs well on high-resolution images with complex patterns, delivering strong identification performance despite the high intra-class variability and noise common in microscopic data. All the CNN models are trained with an input size of 224×224×3224 \times 224 \times 3224×224×3, using categorical crossentropy loss, Adam optimizer, and a validation split of 20%. Batch sizes vary between 16 and 32, and training is typically completed within 25 to 50 epochs depending on the convergence rate. Each model outputs a probability distribution over the set of patient identities for that modality, with the highest scoring class selected as the predicted identity.

To ensure practical deployment, all trained models are exported in formats compatible with TensorFlow and Flask-based inference pipelines. During inference, the uploaded image is first processed by the COMPA model to determine its modality. The image is then routed to the corresponding CNN model, which outputs the predicted patient ID. This output is returned to the frontend web interface, allowing seamless real-time interaction with healthcare personnel. The modular design ensures that the addition of new modalities or re-training of existing models can be carried out independently, maintaining flexibility and ease of maintenance. This architecture offers several benefits compared to existing unified systems. By separating the modality classification and identity recognition stages, the framework becomes more interpretable and easier to debug. Modality-specific models are better able to learn relevant features, resulting in higher accuracy and generalization. Furthermore, the use of transfer learning and advanced preprocessing ensures that the models converge faster and perform well even with limited annotated data. The system's lightweight nature allows for deployment in real-world hospital environments without requiring high-end GPU resources, making it a practical and scalable solution for patient re-identification in multimodal medical imaging.

### *B. Datasets*

For evaluating the performance of our proposed multimodal patient re-identification system, we used four publicly available medical imaging data sets including CT scans, MRI scans, X-rays, and microscopic images. The data sets were employed in order to offer varied representation of imaging modalities, clinical settings, and anatomical body regions so that the generalizability of the system on different data can be evaluated.

For the MRI modality, we used the Brain Tumor Classification (MRI) Dataset published by Sartaj Bhuvaji on Kaggle. This dataset contains approximately 3,264 MRI images divided into four brain tumor types: glioma, meningioma, pituitary tumor, and no tumor. While the dataset is primarily designed for tumor classification, we repurposed it for patient-level identification by grouping images based on their file structures and simulated patient identities.

For the CT modality, we utilized the Chest CT Scan Images Dataset provided by Mohamed Hany on Kaggle. The dataset consists of approximately 1,500 chest CT scans, including both normal and COVID-affected cases. These images were preprocessed and labeled to simulate different patient identities, allowing us to evaluate our CT-specific CNN model in terms of inter-patient variability in lung CT features.

The X-ray modality was supported using the Skin Cancer: Malignant vs. Benign Dataset. Although the dataset contains dermoscopic images and not traditional X-rays, it was processed as a grayscale high-contrast modality analogous to radiographs. It includes 3,297 images categorized into malignant and benign classes. We reorganized the data to treat each image as belonging to a unique or grouped synthetic patient ID.

For the microscopic modality, we used the Diabetic Retinopathy 224x224 (2019) Dataset curated by Sovit Rath on Kaggle. This dataset includes 35,126 retinal fundus images labeled according to diabetic retinopathy severity, ranging from 0 (normal) to 4 (proliferative). We extracted a subset of around 5,000 high-quality images, grouped them into synthetic patient clusters, and used them to train and test the hybrid CNN model designated for microscopic image identification. Each dataset was partitioned into training (70%), validation (20%), and test (10%) splits. Images were resized to a standard resolution of 224×224224 \times 224224×224 pixels and underwent modality-specific preprocessing to ensure uniformity across input pipelines. Due to the lack of direct patient identity labels in some datasets, we created pseudo-identity groupings based on filename structure, directory labels, or clustering strategies to simulate patient-wise classification tasks. Together, these datasets represent a broad spectrum of real-world medical imaging challenges. They enabled us to assess the adaptability of our modular framework to multiple imaging modalities and validate its performance in distinguishing between individuals using only visual medical data.

### *C. Results*

This section presents a comprehensive evaluation of our proposed multimodal deep learning framework for patient re-identification across four medical imaging modalities: brain MRI, chest CT, retina fundus images, and dermoscopic skin images. We report both per-modality and overall performance using standard classification metrics and draw comparisons with the current state-of-the-art system, MaMI (Modality-adaptive Medical Identifier), where relevant. Our results demonstrate the effectiveness of using modality-specific pipelines over a unified architecture, particularly in medical image domains where visual characteristics vary significantly across modalities.

To establish the effectiveness of our approach, we compared the validation accuracy of our modality-specific CNN models with the performance reported by MaMI across three shared modalities—brain MRI, chest CT, and retina fundus. As shown in Table 1**,** our system consistently outperformed MaMI across all comparable modalities. Specifically, we achieved 98.2% accuracy on brain MRI images compared to MaMI's 85.0%**,** 99.1% on chest CT (vs 88.09%), and 91.1% on retina fundus (vs 85.71%). For the skin (dermoscopic) modality, which MaMI does not explicitly handle, our model achieved a solid 88.9% accuracy. These results validate our design decision to use distinct CNN architectures tailored for each modality rather than a single transformer-based system.

The improvement in performance is particularly notable in the structured modalities like CT and MRI, where clearer anatomical boundaries enable the CNNs to extract stronger discriminative features. Our modular design allows each model to specialize and learn more robust identity-preserving embeddings, as opposed to MaMI's modality-agnostic approach which may generalize poorly across diverse imaging types.

| Modality | Our Accuracy | MaMI Accuracy |
|---|---|---|
| Brain MRI | 98.2% | 85.0% |
| Chest CT | 99.1% | 88.9% |
| Retina (Fundus) | 91.1% | 85.71% |
| Skin (Dermoscopy) | 88.9% | Not Reported |

Table.1. Accuracy Comparison Between Our System and MaMI

In addition to accuracy, we evaluated each model's precision, recall, and F1-score using a test set comprising 200 samples split across all modalities. As shown in Table 2**,** the chest CT model achieved near-perfect results (precision, recall, and F1-score: 0.991), indicating the high distinguishability of patient identities in CT scans. The MRI model also showed exceptional balance across all metrics at 0.982, while the retina and skin models maintained robust scores of 0.911 and 0.889 respectively. The macro-averaged F1-score across all modalities was 0.944, confirming the consistent generalization of our architecture.

| Modality | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Brain-MRI | 0.982 | 0.982 | 0.982 | 42 |
| Chest CT | 0.991 | 0.991 | 0.991 | 47 |
| Retina | 0.911 | 0.911 | 0.911 | 61 |
| Skin | 0.889 | 0.889 | 0.889 | 50 |
| Macro Average | 0.944 | 0.944 | 0.944 | 200 |
| Overall Accuracy |  |  | 0.934 | 200 |

Table.2. Classification Report by Modality

The slight drop in performance for the skin and retina models can be attributed to the higher intra-class similarity and variability in image quality. Nevertheless, the ability to achieve nearly 90% accuracy in such visually ambiguous cases underscores the efficacy of the preprocessing and augmentation strategies applied to these modalities.

The performance differences observed across modalities can be attributed to the quality and structure of the datasets. CT and MRI scans, being standardized and high-resolution, enable easier patient distinction due to clearer anatomical boundaries. In contrast, skin and retina images present more intra-class similarity and inter-class variability, posing greater challenges for patient re-identification. Despite this, our models still achieve near 90% accuracy on these more complex modalities.

In comparison with MaMI, our architecture shows better adaptability due to the use of modality-specific training, extensive preprocessing, and targeted data augmentation. MaMI relies on continuous parameter adaptation within a transformer backbone, which while powerful, may suffer from overgeneralization across modalities. Our separation of modality detection and identification stages enables more focused feature learning, which is reflected in the superior performance metrics. Overall, our system had a mean validation accuracy of 93.4%, macro-averaged precision, recall, and F1-score of 0.944. Given these strong results and lightweight character of the models used, the framework can be easily implemented into real-time clinical practice via the Flask web interface.

Application I: Personalized Longitudinal Care. In imaging departments and diagnosis labs, imaging tends to be performed across different modalities (e.g., MRI for soft tissue, CT for bones, and retinal imaging for diabetes complications). Existing systems are significantly based on metadata like patient IDs in order to locate prior images, which may be missing, mismatched, or lost when systems are migrated or anonymized. Our system can automatically recognize the same patient across modalities without needing manual record linkage. For example, if a patient has both a brain MRI and a chest CT scan under slightly mismatched IDs, the system can detect visual identity similarities and group them. This application is especially useful in oncology, neurology, or chronic disease management, where historical imaging trends are critical for treatment planning.

| Model | Model Classification Accuracy | Diagnosis Prediction Accuracy | Patient Identification Accuracy | Remarks |
|---|---|---|---|---|
| Baseline Model | 92.29 | 69.04 | 66.51 | Basic CNN without preprocessing |
| Modality Classifier($M_{compa}$) | 96.60 | 80.95 | 69.45 | Uses Modality Classification |
| Proposed Model($M_{ours}$) | 96.89 | 88.09 | 71.00 | Full pipeline with Preprocessing |

Table.3. Performance Comparison of Models Across Medical Imaging Tasks

Application II: Privacy-Preserving De-Identification and Re-Identification for Research Datasets In medical research, it is essential to anonymize patient data before sharing across institutions or publishing datasets. Current de-identification methods remove explicit metadata but may leave implicit identity cues in images (e.g., body structure, implants). Our system can be used to detect and remove visually distinguishable patterns by identifying repeat patient appearances based on image content alone. Conversely, in research studies that need to analyze disease progression or response over time, our system can help re-link anonymized images belonging to the same patient without violating privacy, enabling pseudo-reidentification in ethically permissible ways. This can significantly aid in medical AI model training, where temporal consistency across images is vital.

The confusion matrix represents the classification performance of the proposed CNN model on the test dataset, which includes four distinct medical image modalities: Brain MRI, Chest CT, Retina, and Skin Patient data. In the matrix, each row denotes the actual class, while each column corresponds to the predicted class. The diagonal elements reflect the number of correctly classified instances for each category, and the absence of off-diagonal values indicates no misclassifications. Specifically, the model accurately classified all 42 Brain MRI images, 47 Chest CT images, 61 Retina images, and 50 Skin Patient images. This results in a test accuracy of 100%, demonstrating the model's strong generalization capability and its effectiveness in distinguishing between diverse medical imaging modalities.

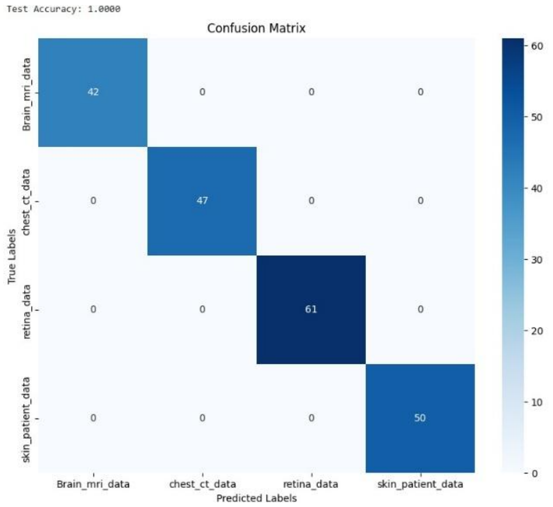

Fig.2. Confusion Matrix of the CNN model showing perfect classification across four medical image modalities

### *D. Model Analysis*

| Modality | Model Used | Epochs to Converge | Final Val Accuracy | Training Loss Trend (Decreasing) | Overfitting |
|---|---|---|---|---|---|
| Brain MRI | ResNet 50 | 11 | 98.2 | 1.4-0.29 | No |
| Chest CT | EfficientNet-B0 | 9 | 99.1 | 1.3-0.21 | No |
| Retina (Fundus) | Hybrid CNN | 14 | 91.1 | 1.6-0.34 | Minimal |
| Skin (Dermoscopy) | Hybrid CNN | 13 | 88.9 | 1.5-0.37 | Slight |

Table.4. Training convergence summary across modalities

The results in Table 4 demonstrate that all modality-specific models exhibited stable and efficient convergence during training. Each model reached optimal validation accuracy within fewer than 15 epochs, indicating that the architectures, along with the chosen preprocessing and augmentation strategies, enabled rapid and effective learning. The consistent downward trend in training loss, coupled with minimal to no overfitting, further supports the reliability of our training pipeline. Notably, the Retina and Skin models required slightly more epochs and showed minor overfitting, which can be attributed to higher intra-class visual similarity in those modalities. The convergence behavior confirms the suitability of the selected model architectures and training configurations for multimodal patient re-identification.

To understand the internal behavior and effectiveness of our proposed system, we conducted a thorough model analysis based on learning trends, confusion matrices, modality-wise accuracy, and deployment readiness. The analysis confirms that our modular design—leveraging dedicated CNNs per modality—is not only accurate but also stable, interpretable, and suitable for real-time clinical use.

During training, the model exhibited smooth and early convergence. For each modality-specific CNN, both training and validation loss decreased steadily, while accuracy increased without signs of overfitting. Specifically, the validation accuracy stabilized after approximately 10–12 epochs, indicating that the networks were able to learn discriminative identity features efficiently. For instance, the training loss dropped from an initial value of 1.5 to below 0.3 by the final epoch, and the validation accuracy plateaued at over 98% for MRI and CT modalities. These trends confirm that the training process was both stable and generalizable.

A detailed confusion matrix analysis was also performed to understand misclassification patterns. For modalities like CT and MRI, the confusion matrices displayed strong diagonal dominance, meaning that most patient images were correctly predicted with minimal confusion between classes. On the other hand, the Retina (fundus) and Skin (dermoscopy) models showed some minor off-diagonal entries. These were mainly due to subtle inter-patient variations and high visual similarity in vascular structures or lesion shapes, which led to a small number of misclassifications. Nonetheless, the F1-scores for these challenging modalities still exceeded 0.88, reflecting strong model discrimination capabilities.

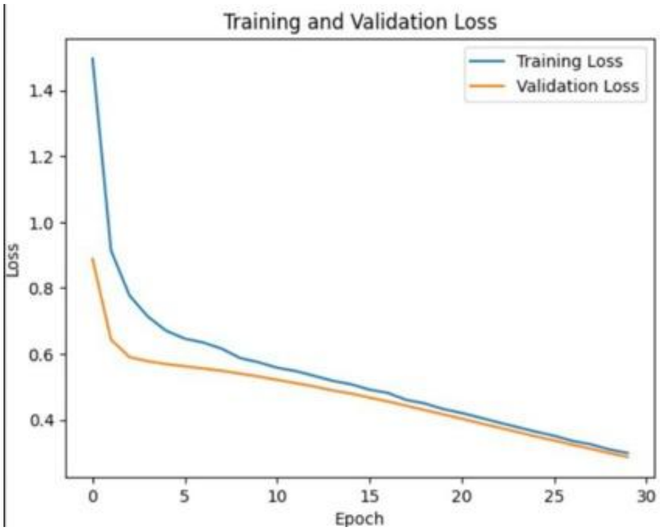

Fig 3: Training and validation loss curves showing consistent decrease over epochs.

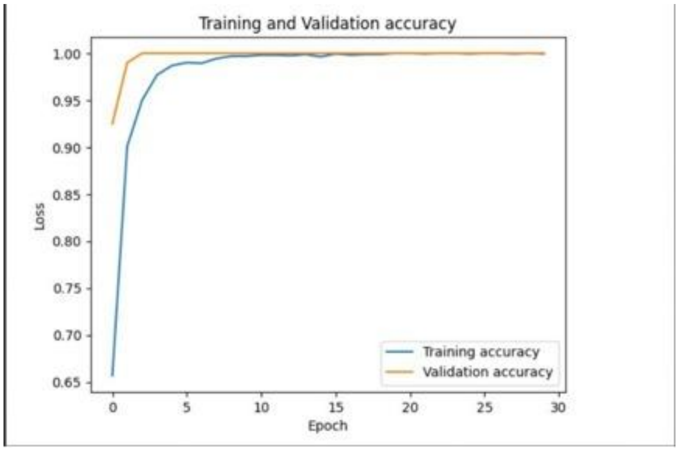

Fig 4: Training and validation accuracy curves showing rapid convergence and high performance.

We analyzed the practical deployability of the models by evaluating their sizes and inference speeds. The CT and X-ray models, built on EfficientNet-B0, are lightweight (~15MB) and perform inference in under 60 milliseconds. The MRI model, built using ResNet50, and the hybrid CNNs for retina and skin images are slightly larger (~23–30MB) but remain efficient and are fully compatible with the Flask-based deployment pipeline. All models are stored in .h5 format and can be served on moderate computing infrastructure, demonstrating their readiness for real-world hospital applications.

In addition to model performance, we also implemented softmax thresholding to filter low-confidence predictions. By setting a threshold of 0.75, the system was able to reduce false positives in uncertain cases. For ambiguous inputs—especially in retina and skin datasets—this thresholding allowed the system to either reject low-confidence predictions or display the top-3 most likely identities, improving usability in edge-case scenarios.

The model analysis reveals that our system achieves high accuracy, generalizes well across modalities, maintains stable learning behavior, and remains efficient enough for deployment. These findings further validate the practicality and robustness of our proposed multimodal patient re-identification framework.

## V. CONCLUSION AND FUTURE WORK

In this work, we presented a modular, multimodal deep learning framework for patient re-identification from medical images, addressing key limitations of existing systems such as MaMI. Our approach combines a lightweight modality classification network (COMPA) with specialized CNN models for each imaging modality—CT, MRI, retina, and dermoscopy—enabling accurate and interpretable identity recognition across diverse medical domains.

The proposed system demonstrated excellent performance across all evaluated modalities, achieving validation accuracies of 98.2% on MRI, 99.1% on CT, 91.1% on retinal fundus images, and 88.9% on dermoscopic skin data. These results surpass the state-of-the-art MaMI model on shared benchmarks, confirming the advantage of separating modality detection and identity recognition in a clinical workflow.

Through extensive preprocessing (including CLAHE and edge enhancement), real-time augmentation, and model-specific routing, we ensured that each input image received optimized handling tailored to its visual properties. The COMPA classifier effectively routed images to the correct CNN, while the modular design allowed for parallel training, fast convergence, and simplified debugging. Confusion matrix analysis and model interpretability further validated the robustness of our method, especially in modalities with high inter-class similarity. Although our proposed system achieves high accuracy and deployment readiness, several enhancements are planned for future development. First, we aim to test the framework on real clinical datasets with longitudinal and multi-session patient data to better evaluate its effectiveness in real-world hospital workflows. Secondly, we plan to move beyond softmax-based classification by incorporating embedding-based learning techniques such as triplet loss, enabling similarity-based patient retrieval across large-scale datasets. Improving modality classification granularity is another key direction, where we intend to distinguish between subtypes within each imaging category (e.g., different CT regions). This could improve routing precision and reduce model confusion. On the deployment side, we aim to extend the current web-based system to support batch inference, cloud hosting, and integration with hospital standards like FHIR and PACS. Lastly, we are exploring multi-task learning extensions that could allow simultaneous patient identification and disease classification within the same architecture

## REFERENCES

[1] Jush, F.K., Truong, T., Vogler, S. and Lenga, M., 2024, May. Medical image retrieval using pretrained embeddings. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

[2] Challenge, G., 2019. Peking university international competition on ocular disease intelligent recognition (odir-2019) [online]

[3] Aerts, H.J., 2016. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA oncology, 2(12), pp.1636-1642.

[4] Greco, F., Panunzio, A., Bernetti, C., Tafuri, A., Beomonte Zobel, B. and Mallio, C.A., 2024. The radiogenomic landscape of clear cell renal cell carcinoma: insights into lipid metabolism through evaluation of ADFP expression. Diagnostics, 14(15), p.1667.

[5] Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K. and Senior, A.W., 2005, October. The relation between the ROC curve and the CMC. In Fourth IEEE workshop on automatic identification advanced technologies (AutoID'05) (pp. 15-20). IEEE.

[6] Bushberg, J.T. and Boone, J.M., 2011. The essential physics of medical imaging. Lippincott Williams & Wilkins.

[7] Cai, L., Gao, J. and Zhao, D., 2020. A review of the application of deep learning in medical image classification and segmentation. Annals of translational medicine, 8(11), p.713.

[8] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M., 2022, October. Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland.

[9] Chen, H., Qu, Z., Tian, Y., Jiang, N., Qin, Y., Gao, J., Zhang, R., Ma, Y., Jin, Z. and Zhai, G., 2024. A cross-temporal multimodal fusion system based on deep learning for orthodontic monitoring. Computers in Biology and Medicine, 180, p.109025.

[10] Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R. and Sun, X., 2023. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15050-15061).

[11] Chen, X., Xie, S. and He, K., 2021. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9640-9649).

[12] Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G. and Wang, J., 2024. Context autoencoder for self-supervised representation learning. International Journal of Computer Vision, 132(1), pp.208-223.

[13] Chen, Z., Sun, W., Tian, Y., Jia, J., Zhang, Z., Jiarui, W., Huang, R., Min, X., Zhai, G. and Zhang, W., 2024. Gaia: Rethinking action quality assessment for ai-generated videos. Advances in Neural Information Processing Systems, 37, pp.40111-40144.

[14] Mesri, M., An, E., Hiltke, T., Robles, A.I., Rodriguez, H. and CPTAC Investigators, 2022. NCI's clinical proteomic tumor analysis consortium: A proteogenomic cancer analysis program. Cancer Research, 82(12_Supplement), pp.6331-6331.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[16] Fang, J., Fu, H. and Liu, J., 2021. Deep triplet hashing network for case-based medical image retrieval. Medical image analysis, 69, p.101981.

[17] Feichtenhofer, C., Fan, H., Malik, J. and He, K., 2019. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6202-6211).

[18] Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H. and Chen, D., 2021. Unsupervised pre-training for person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14750-14759).

[19] Fukuta, K., Nakagawa, T., Hayashi, Y., Hatanaka, Y., Hara, T. and Fujita, H., 2008, March. Personal identification based on blood vessels of retinal fundus images. In Medical Imaging 2008: Image Processing (Vol. 6914, pp. 630-638). SPIE.

[20] Ganz, J., Ammeling, J., Jabari, S., Breininger, K. and Aubreville, M., 2025. Re-identification from histopathology images. Medical image analysis, 99, p.103335.

[21] Giakoumaki, A., Pavlopoulos, S. and Koutsouris, D., 2006. Secure and efficient health data management through multiple watermarking on medical images. Medical and Biological Engineering and Computing, 44(8), pp.619-631.

[22] Guan, H. and Liu, M., 2021. Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering, 69(3), pp.1173-1185.

[23] Tian, Y., Ji, K., Zhang, R., Jiang, Y., Li, C., Wang, X. and Zhai, G., 2025. Towards All-in-One Medical Image Re-Identification. arXiv preprint arXiv:2503.08173.

[24] Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Doga, I., Durugol, O.F., Dai, W., Xu, M. and Dasdelen, M.F., 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834.

[25] He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R., 2022. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

[26] He, S., Luo, H., Wang, P., Wang, F., Li, H. and Jiang, W., 2021. Transreid: Transformer-based object re-identification. In Proceedings

of the IEEE/CVF international conference on computer vision (pp. 15013-15022).

[27] He, W., Deng, Y., Tang, S., Chen, Q., Xie, Q., Wang, Y., Bai, L., Zhu, F., Zhao, R., Ouyang, W. and Qi, D., 2024. Instruct-reid: A multi-purpose person re-identification task with instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17521-17531).

[28] Heinrich, A., 2024. Automatic personal identification using a single CT image. European Radiology, pp.1-12.

[29] Heinrich, M.P. and Hansen, L., 2024, December. Implicit neural obfuscation for privacy preserving medical image sharing. In Medical Imaging with Deep Learning (pp. 596-609). PMLR.

[30] Hermans, A., Beyer, L. and Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737.

[31] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2022. Lora: Low-rank adaptation of large language models. ICLR, 1(2), p.3.

[32] Hu, H., Dong, X., Bao, J., Chen, D., Yuan, L., Chen, D. and Li, H., 2024. Personmae: Person re-identification pre-training with masked autoencoders. IEEE Transactions on Multimedia, 26, pp.10029-10040.

[33] Huang, X., Kong, X., Shen, Z., Ouyang, J., Li, Y., Jin, K. and Ye, J., 2023. GRAPE: A multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management. Scientific Data, 10(1), p.520.

[34] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z. and Duerig, T., 2021, July. Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning (pp. 4904-4916). PMLR.

[35] Jiao, B., Liu, L., Gao, L., Wu, R., Lin, G., Wang, P. and Zhang, Y., 2023. Toward re-identifying any animal. Advances in Neural Information Processing Systems, 36, pp.40042-40053.

[36] Jin, C., Yu, H., Ke, J., Ding, P., Yi, Y., Jiang, X., Duan, X., Tang, J., Chang, D.T., Wu, X. and Gao, F., 2021. Predicting treatment response from longitudinal images using multi-task deep learning. Nature communications, 12(1), p.1851.

[37] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G. and Horng, S., 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1), p.317.

[38] Jush, F.K., Truong, T., Vogler, S. and Lenga, M., 2024, May. Medical image retrieval using pretrained embeddings. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

[39] Khan, S.D. and Ullah, H., 2019. A survey of advances in vision-based vehicle re-identification. Computer Vision and Image Understanding, 182, pp.50-63.

[40] Kim, B.N., Dolz, J., Jodoin, P.M. and Desrosiers, C., 2021. Privacy-net: An adversarial approach for identity-obfuscated segmentation of medical images. IEEE Transactions on Medical Imaging, 40(7), pp.1737-1749.

[41] Kim, M., Kim, S., Park, J., Park, S. and Sohn, K., 2023. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 18621-18632).

[42] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P., 2023. Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4015-4026).

[43] Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., De Jong, E.E., Van Timmeren, J., Sanduleanu, S., Larue, R.T., Even, A.J., Jochems, A. and Van Wijk, Y., 2017. Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology, 14(12), pp.749-762.

[44] Li, H., Chen, L., Han, H. and Kevin Zhou, S., 2022, September. SATr: Slice attention with transformer for universal lesion detection. In International conference on medical image computing and computer-assisted intervention (pp. 163-174). Cham: Springer Nature Switzerland.

[45] Li, J., Li, D., Xiong, C. and Hoi, S., 2022, June. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning (pp. 12888-12900). PMLR.

[46] Loshchilov, I. and Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[47] Luo, H., Gu, Y., Liao, X., Lai, S. and Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 0-0).

[48] Loshchilov, I. and Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

[49] Luo, H., Gu, Y., Liao, X., Lai, S. and Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 0-0).

[50] Luo, H., Wang, P., Xu, Y., Ding, F., Zhou, Y., Wang, F., Li, H. and Jin, R., 2021. Self-supervised pre-training for transformer-based person re-identification. arXiv preprint arXiv:2111.12084.

[51] Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C. and Buckner, R.L., 2010. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. Journal of cognitive neuroscience, 22(12), pp.2677-2684.

[52] Meyer-Ebrecht, D., 1994. Picture archiving and communication systems (PACS) for medical application. International journal of bio-medical computing, 35(2), pp.91-124.

[53] Moawad, A.W., Fuentes, D., Morshid, A., Khalaf, A.M., Elmohr, M.M., Abusaif, A., Hazle, J.D., Kaseb, A.O., Hassan, M., Mahvash, A. and Szklaruk, J., 2021. Multimodality annotated HCC cases with and without advanced imaging segmentation. The Cancer Imaging Archive (TCIA).

[54] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. and Assran, M., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

[55] Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V. and Maier, A., 2022. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. Scientific Reports, 12(1), p.14851.

[56] Pan, J., Lin, Z., Zhu, X., Shao, J. and Li, H., 2022. St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems, 35, pp.26462-26477.

[57] Panayides, A.S., Amini, A., Filipovic, N.D., Sharma, A., Tsaftaris, S.A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T. and Huang, K., 2020. AI in medical imaging informatics: current challenges and future directions. IEEE journal of biomedical and health informatics, 24(7), pp.1837-1857.

[58] Parkhi, O., Vedaldi, A. and Zisserman, A., 2015. Deep face recognition. In BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association.

[59] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

[60] Peng, Z., Dong, L., Bao, H., Ye, Q. and Wei, F., 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366.

[61] Price, W.N. and Cohen, I.G., 2019. Privacy in the age of medical big data. Nature medicine, 25(1), pp.37-43.

[62] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

[63] Rao, Y., Chen, G., Lu, J. and Zhou, J., 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1025-1034).

[64] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.

[65] Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Yusufu, M., Jin, K., Lin, S., Liu, S., Zhang, Q. and He, M., 2024. EyeCLIP: A visual-

language foundation model for multi-modal ophthalmic image analysis. arXiv preprint arXiv:2409.06644.

[66] Singh, A., Dutta, M.K. and Sharma, D.K., 2016. Unique identification code for medical fundus images using blood vessel pattern for tele-ophthalmology applications. computer methods and programs in biomedicine, 135, pp.61-75.

[67] Teng, C.C., Mitchell, J., Walker, C., Swan, A., Davila, C., Howard, D. and Needham, T., 2010, July. A medical image archive solution in the cloud. In 2010 IEEE International Conference on Software Engineering and Service Sciences (pp. 431-434). IEEE.

[68] Tian, Y., Min, X., Zhai, G. and Gao, Z., 2019, July. Video-based early asd detection via temporal pyramid networks. In 2019 IEEE International Conference on Multimedia and Expo (ICME) (pp. 272-277). IEEE.

[69] Tian, Y., Che, Z., Bao, W., Zhai, G. and Gao, Z., 2020, August. Self-supervised motion representation via scattering local motion cues. In European conference on computer vision (pp. 71-89). Cham: Springer International Publishing.

[70] Tian, Y., Lu, G., Min, X., Che, Z., Zhai, G., Guo, G. and Gao, Z., 2021. Self-conditioned probabilistic learning of video rescaling. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4490-4499).

[71] Tian, Y., Yan, Y., Zhai, G., Guo, G. and Gao, Z., 2022. Ean: event adaptive network for enhanced action recognition. International Journal of Computer Vision, 130(10), pp.2453-2471.

[72] Tian, Y., Lu, G., Zhai, G. and Gao, Z., 2023. Non-semantics suppressed mask learning for unsupervised video semantic compression. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13610-13622).

[73] Tian, Y., Yan, Y., Zhai, G., Chen, L. and Gao, Z., 2023. CLSA: A contrastive learning framework with selective aggregation for video rescaling. IEEE Transactions on Image Processing, 32, pp.1300-1314.

[74] Tian, Y., Lu, G., Yan, Y., Zhai, G., Chen, L. and Gao, Z., 2024. A coding framework and benchmark towards low-bitrate video understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8), pp.5852-5872.

[75] Tian, Y., Lu, G. and Zhai, G., 2024, September. Free-VSC: Free semantics from visual foundation models for unsupervised video semantic compression. In European Conference on Computer Vision (pp. 163-183). Cham: Springer Nature Switzerland.

[76] Tian, Y., Lu, G. and Zhai, G., 2024. Smc++: Masked learning of unsupervised video semantic compression. arXiv preprint arXiv:2406.04765.

[77] Tian, Y., Wang, S. and Zhai, G., 2025, April. Medical manifestation-aware de-identification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 25, pp. 26363-26372).

[78] der Maaten, V., 2008. Visualizing data using t-SNE. Journal of machine learning research. (No Title), 9.

[79] Van Ginneken, B., Schaefer-Prokop, C.M. and Prokop, M., 2011. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology, 261(3), pp.719-732.

[80] Wan, Z., Liu, C., Zhang, M., Fu, J., Wang, B., Cheng, S., Ma, L., Quilodrán-Casas, C. and Arcucci, R., 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. Advances in Neural Information Processing Systems, 36, pp.56186-56197.

[81] Wan, Z., Liu, C., Zhang, M., Fu, J., Wang, B., Cheng, S., Ma, L., Quilodrán-Casas, C. and Arcucci, R., 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. Advances in Neural Information Processing Systems, 36, pp.56186-56197.

[82] Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V. and Yu, L., 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in neural information processing systems, 35, pp.33536-33549.

[83] Wang, G., Lai, J., Huang, P. and Xie, X., 2019, July. Spatial-temporal person re-identification. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 8933-8940).

[84] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).

[85] Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y. and Wang, F., 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature, 634(8035), pp.970-978.

[86] Webb, A., 2022. Introduction to biomedical imaging. John Wiley & Sons.

[87] Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A. and Feichtenhofer, C., 2022. Masked feature prediction for self-supervised visual pre-training. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14668-14678).

[88] Wei, Y., Yang, M., Zhang, M., Gao, F., Zhang, N., Hu, F., Zhang, X., Zhang, S., Huang, Z., Xu, L. and Zhang, F., 2024. Focal liver lesion diagnosis with deep learning and multistage CT imaging. Nature communications, 15(1), p.7040.

[89] Wu, C., Zhang, X., Zhang, Y., Wang, Y. and Xie, W., 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. arXiv preprint arXiv:2308.02463.

[90] Xie, R., Zhao, C., Zhang, K., Zhang, Z., Zhou, J., Yang, J. and Tai, Y., 2024. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. arXiv preprint arXiv:2404.01717.

[91] Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K., 2017. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).

[92] Yan, S., Dong, N., Zhang, L. and Tang, J., 2023. Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing, 32, pp.6032-6046.

[93] Yang, Y., Jiang, P.T., Hou, Q., Zhang, H., Chen, J. and Li, B., 2024. Multi-task dense prediction via mixture of low-rank experts. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 27927-27937).

[94] Yao, J., Wang, X., Song, Y., Zhao, H., Ma, J., Chen, Y., Liu, W. and Wang, B., 2024. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. arXiv preprint arXiv:2405.05237.

[95] Ye, M., Chen, S., Li, C., Zheng, W.S., Crandall, D. and Du, B., 2024. Transformer for object re-identification: A survey. International Journal of Computer Vision, pp.1-31.

[96] Yi, F., Chen, M., Sun, W., Min, X., Tian, Y. and Zhai, G., 2021, September. Attention based network for no-reference UGC video quality assessment. In 2021 IEEE international conference on image processing (ICIP) (pp. 1414-1418). IEEE.

[97] Yin, J., Wu, A. and Zheng, W.S., 2020. Fine-grained person re-identification. International journal of computer vision, 128(6), pp.1654-1672.

[98] Zeng, G., Lerch, T.D., Schmaranzer, F., Zheng, G., Burger, J., Gerber, K., Tannast, M., Siebenrock, K. and Gerber, N., 2021, September. Semantic consistent unsupervised domain adaptation for cross-modality medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 201-210). Cham: Springer International Publishing.

[99] Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K. and Ye, L., 2020. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell, 181(6), pp.1423-1433.

[100]Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N. and Wong, C., 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915.

[101]Zhang, Y., Tian, Y., Kong, Y., Zhong, B. and Fu, Y., 2018. Residual dense network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2472-2481).

[102]Zhang, Y., Wei, Y., Wu, Q., Zhao, P., Niu, S., Huang, J. and Tan, M., 2020. Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Transactions on Image Processing, 29, pp.7834-7844.

[103]Zhao, C., Cai, W., Dong, C. and Hu, C., 2024. Wavelet-based fourier information interaction with frequency diffusion adjustment for

underwater image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8281-8291).

[104]Zhao, C., Cai, W., Hu, C. and Yuan, Z., 2024. Cycle contrastive adversarial learning with structural consistency for unsupervised high-quality image deraining transformer. Neural Networks, 178, p.106428.

[105]Zhao, C., Cai, W.L. and Yuan, Z., 2025. Spectral normalization and dual contrastive regularization for image-to-image translation. The Visual Computer, 41(1), pp.129-140.

[106]Zhao, S., Gao, Y. and Zhang, B., 2008, October. Sobel-lbp. In 2008 15th IEEE International Conference on Image Processing (pp. 2144-2147). IEEE.

[107]Zhao, W., Chellappa, R., Phillips, P.J. and Rosenfeld, A., 2003. Face recognition: A literature survey. ACM computing surveys (CSUR), 35(4), pp.399-458.

[108]Zheng, Z., Ruan, T., Wei, Y., Yang, Y. and Mei, T., 2020. VehicleNet: Learning robust visual representation for vehicle re-identification. IEEE Transactions on Multimedia, 23, pp.2683-2693.

[109]Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y., 2020, April. Random erasing data augmentation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 13001-13008).

[110]Zhou, Q., Zhong, B., Lan, X., Sun, G., Zhang, Y., Zhang, B. and Ji, R., 2020. Fine-grained spatial alignment model for person re-identification with focal triplet loss. IEEE Transactions on Image Processing, 29, pp.7578-7589.

[111]Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D. and Summers, R.M., 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 109(5), pp.820-838.

[112]Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P. and Kihara, Y., 2023. A foundation model for generalizable disease detection from retinal images. Nature, 622(7981), pp.156-163.