IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 January 2025, accepted 29 January 2025, date of publication 4 February 2025, date of current version 12 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3538325

RESEARCH ARTICLE

DuCo-Net: Dual-Contrastive Learning Network for Medical Report Retrieval Leveraging Enhanced Encoders and Augmentations

ZAHID UR RAHMAN[1], JU-HWAN LEE[1], DANG THANH VU[2], IQBAL MURTZA[1,3], AND JIN-YOUNG KIM[1], (Member, IEEE)

[1]Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, South Korea
[2]Research Center, AISeed Inc., Gwangju 61186, South Korea
[3]Department of Creative Technologies, Faculty of Computing and AI, Air University Islamabad, Islamabad 44000, Pakistan

Corresponding author: Jin-Young Kim (beyondi@jnu.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) (Artificial Intelligence Innovation Hub, 50) under Grant RS-2021-II212068, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization Program grant funded by the Korea Government (MSIT) under Grant IITP-2025-RS-2022-00156287, 50.

ABSTRACT The conventional process of generating medical radiology reports is labor-intensive and time-consuming, requiring radiologists to describe findings meticulously from imaging studies. This manual approach often causes undesirable delays in patient care. Despite advancements in computer vision and deep learning, developing an effective computer-aided solution to generate automated medical reports remains challenging. The recent advancements in deep learning technology, especially with the advent of contrastive learning, have shown significant performance in natural language supervision. However, their application to medical report generation, particularly in the domain of chest x-rays (CXR), has been limited due to the lack of large annotated datasets. Many studies have proposed multimodal contrastive learning schemes to address the data scarcity problem for natural images. However, none of these techniques have been efficiently explored in terms of medical report generation. This study addresses these challenges by proposing a dual contrastive learning network (DuCo-Net) containing backbone and augmented networks. The backbone network is trained on the original data, while the augmented network emphasizes cross-model augmentation learning in a unified framework. DuCo-Net enables two complementary learning mechanisms: intra-modal learning, where each network learns specialized features within its modality (either image or text), and inter-modal learning, which captures relationships between image and text modalities through a combined loss function. This dual learning approach leverages modified DenseNet121 and BioBERT models with advanced pooling techniques specifically tailored for handling medical data. Comprehensive evaluations on two publicly available datasets demonstrate that DuCo-Net significantly outperforms current benchmarks. On the Indiana University Chest X-rays dataset, our proposed methodology demonstrates significant improvements across standard metrics (BLEU-1: 0.50, ROUGE: 0.40, METEOR: 0.24, F1: 0.40). For the MIMIC-CXR dataset, the framework maintains robust performance (BLEU-1: 0.42, ROUGE: 0.34, METEOR: 0.20, F1: 0.34), representing substantial improvements over existing state-of-the-art approaches in medical report generation.

INDEX TERMS Medical report retrieval, contrastive learning, multi-modal learning, deep learning, chest x-rays, medical image augmentation, radiology.

I. INTRODUCTION

According to the World Health Organization, more than three and a half billion medical diagnosis examinations

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

are conducted worldwide yearly, such as x-rays [1]. These examinations are manually observed by radiologists for the radiology report, a written document that provides detailed information about a patient's medical history, symptoms, and the result of relevant radiological exams. The report is typically divided into several sections including comparisons, indications, findings, and impressions as depicted in Fig.1. The findings section is particularly important because it describes the medical observations made by the radiologist during the exam, including any abnormal conditions that may be present. This time-consuming process is the primary reason for unwanted delays in reporting after a radiology scan [2].

Automatic medical report generation using vision language technology is a significant area of research using radiography scan images [3] and significantly affects health care. Computer vision models, when combined with natural language processing (NLP), can be trained to recognize various features in medical images, potentially classifying abnormalities and translating these visual findings into coherent written reports. These models can make the process of screening radiography scans more efficient and reduce the workload of radiologists [4]. Additionally, vision language models can process numerous data to perform mass screening more quickly as compared to manual methods.

However, the complexity of automatically interpreting abnormalities in medical imaging presents significant challenges. Although numerous studies have been conducted on medical report generation using the conventional encoder-decoder architectures following the image captioning paradigm [5], [6], [7], [8], [9], [10], [11], these approaches often yielded to limited performance due to the intricate, lengthy, and biased syntax of medical reports.

Innovative transformer-based multimodal deep learning approaches with robust attention mechanisms have emerged in response to these limitations. These models offer the extraction of pertinent features and have demonstrated promising outcomes in medical imaging-based computer-aided diagnostics [12], [13], [14], [15], [16]. By harnessing the strengths of the visual and textual data, these models may overcome the constraints encountered by traditional deep learning approaches. Nonetheless, the development of methods capable of generating reports containing accurate information to advance patient care remains challenging [17]. One of the key obstacles in most of the proposed vision-based transformer approaches is the scarcity of annotated data, which is important for training such large-scale models effectively [18], [19]. This limitation is particularly acute in the medical domain, where data annotation is time-consuming and labor-intensive.

Recently, contrastive learning has emerged as a strong alternative to traditional supervised learning approaches, offering the ability to leverage unlabeled textual data without explicit supervision [20]. Unlike conventional computer vision models constrained to predefined object categories, contrastive learning can be adapted for various downstream tasks. However, existing contrastive learning approaches, such as contrastive learning image pretraining (CLIP) [21], require massive training data of 400 million image-text pairs, making them highly data-intensive. These approaches may not generalize well to specialized domains, such as medical imaging, primarily due to the scarcity of large-scale data. The specified model is utilized for the medical reports retrieval by employing pre-trained encoders on natural images [22], [23]. However, it is pertinent to note that this approach may not be effective, as models trained on natural data lack generalizability to the medical domain [24]. The significant disparity between natural images and medical images, coupled with the specialized nature of medical reports, presents a substantial obstacle in directly applying these models to medical tasks.

Some of the studies have utilized multi modal approaches integrating contrastive learning techniques for natural image datasets [25], [26], [27], [28], [29]. The results demonstrate that data augmentation and contrastive learning in multi modal frameworks can significantly enhance accuracy by comparing heterogeneous modalities in a common similarity space while reducing the required training data. Moreover, these multi-modal contrastive learning methods have shown notable computational efficiency, achieving faster convergence during training [28]. However, a significant gap remains in applying these techniques to medical report generation. The potential benefits of multi-modal contrastive learning, including improved accuracy with limited data and increased computational efficiency, have yet to be fully explored and leveraged in the context of generating medical reports from imaging data.

To this end, this paper introduces, a novel dual-contrastive learning network (DuCo-Net) designed for efficient medical report retrieval from chest x-rays (CXRs). Inspired by DeCLIP [21], DuCo-Net uses a joint network of two contrastive learning frameworks to preserve high-level similarities in medical image-report pairs by leveraging potential data. The backbone network learns representations from original data, while its augmented counterpart network learns through various augmented forms, helping the backbone network better discriminate between similar pairs by providing controlled variations. This is crucial for medical images where subtle differences can be critical - the augmented network helps the backbone network learn what features truly distinguish one case from another, even when they appear similar. This approach differs from DeCLIP as both networks converge on the original data. Drawing from BYOL [29], most layers in the augmented network are frozen in a strategic manner, ensuring stable training on small-scale datasets and overcoming overfitting. DuCo-Net learns joint representations of images and reports, capturing diverse and crucial semantic relationships in medical data. The model incorporates enhanced versions of DenseNet121 [30] and BioBERT [31], specifically tailored for medical data processing. In addition, DuCo-Net outperforms

state-of-the-art models across natural language generation (NLG) metrics (bilingual evaluation understudy [BLEU], recall-oriented understudy for gisting evaluation [ROUGE], and the metric for evaluation of translation with explicit ordering [METEOR]) and the F1 score on the MIMIC-CXR [2] and Indiana University Chest X-ray Collection (IU-Xray) [32] datasets.

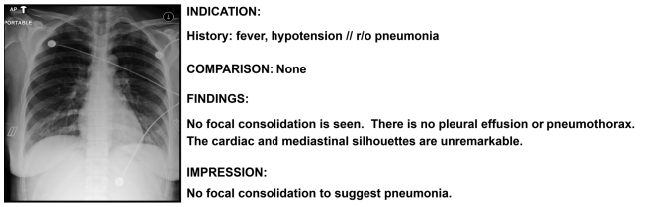


FIGURE 1. Radiology report sample from Mimic-CXR dataset.

The key contributions of this work are as follows:

1) The proposed DuCo-Net architecture employs two contrastive learning schemes jointly trained on original and augmented data, where the backbone network captures robust base representations from original data while the augmented network learns invariant features from augmented versions. These variations help the backbone network better discriminate between similar cases, which is particularly crucial for medical imaging where subtle differences are critical. Unlike single-network contrastive learning which relies on only original positive-negative pairs, our dual-network approach effectively multiplies the learning instances through augmented pairs, making it particularly efficient for small-scale datasets while maintaining training stability through strategic parameter freezing and enhancing representation learning.
2) Modified versions of DenseNet121 and BioBERT models are proposed, as dedicated encoders in the DuCo-Net framework, specifically tailored for the extraction of visual and textual features from medical data.
3) DuCo-Net is validated on two benchmark datasets (MIMIC-CXR and IU-Xray). Extensive experiments demonstrate the state-of-the-art performance of the prosposed model in medical report generation and retrieval tasks using NLG metrics and the F1 score.

The remainder of this paper is organized as follows. Section II reviews traditional and recent methods for medical report generation. Next, Section III introduces the proposed DuCo-Net architecture. Section IV includes an overview and preliminaries. Then, Section V presents the experimental results. Section VI introduces the discussion and limitations of the proposed work. Finally, Section VII summarizes the conclusions.

II. RELATED WORK

The advent of advanced deep learning architectures, particularly the multimodal encoder-decoder frameworks, has revolutionized the field of image captioning [33]. These frameworks integrate computer vision and NLP techniques to generate text from images. The computer vision models predominate in extracting salient visual features, whereas the NLP models adeptly process the textual data and integrate them with visual information for coherent text generation.

A. CNN-RNN ARCHITECTURE

A predominant deep learning paradigm for image captioning jointly employs a convolutional neural network (CNN) [34] and recurrent neural network (RNN) [35]. The CNN is known for its efficacy in image analysis and serves as a visual feature extractor, whereas the RNN is generally used for handling and processing the textual aspects. The CNN-RNN architecture has been widely adopted for medical report generation across numerous studies [36], [37], [38]. In the medical domain, CNNs are used to extract relevant features from radiological images such as CXR and are particularly effective at extracting multiscale features from medical images [39], [40], [41], whereas RNNs are particularly trained to decode these features into comprehensive medical reports concurrently. The use of multimodal deep learning technologies have shown great potential in improving the efficiency of generating accurate medical reports. These approaches save time for healthcare professionals, reduce their workload, and improve the overall consistency and reliability of the reports. However, these models still have significant limitations. A major drawback is their inefficiency in generating complex, lengthy medical reports, primarily due to the vanishing gradient problem of RNNs [42].

Long short-term memory (LSTMs) networks were introduced as an alternative to traditional RNNs to mitigate the problem of vanishing gradients [9], [43]. The LSTM network demonstrates superior performance in handling lengthy medical reports, yielding improved accuracy compared to the conventional RNN models. The LSTM methodology initially identifies and localizes anomalies, extracting them as regions of interest within medical images. Subsequently, these regions are encoded in conjunction with medical reports for LSTM training. Despite these capabilities, LSTM models may still encounter challenges when tasked with generating extensively detailed reports.

B. TRANSFORMER BASED MODELS

The advent of transformer-based models significantly advanced the field of natural language generation, improving the quality and coherence of generated text through their powerful attention-based mechanisms [44]. Several studies have effectively utilized them for medical report generation by replacing RNNs and LSTM newtorks [45], [46], [47]. These models have demonstrated superior performance in processing medical reports, employing their robustness in self-attention mechanisms to maintain contextual awareness across long sequences. Consequently, with a vast dynamic

memory, it can access and prioritize relevant information regardless of its position in the input. Further, these models are faster as compared to the conventional RNNs, and LSTM models by leveraging GPU parallelization. In medical report generation, the problem is not limited to the challenge of generating longer medical reports. Rather, identifying anomalies in the images using semantic properties of textual reports is also essential [48]. Usually in CXR images, the normal regions occupy more space and dominate over the abnormalities inside images. Although CNN-transformer architectures have shown potential, they often struggle to capture the fine-grained details crucial for identifying subtle abnormalities in medical images. In contrast, the vision transformer is proficient at learning hierarchical representations directly from image patches, providing a more considerable approach to feature extraction [49]. To this end, AlignTransformer [50] was introduced by replacing the CNN with a vision transformer including a hierarchical attention mechanism, which first predicts disease tags in images and then uses these disease tags to learn fine-grained visual features. Similarly, a memory-driven transformer with augmented memory blocks was proposed to capture detailed visual features from images using text [51]. Additionally, a medical concepts generation network (MCGN) was included in this model to predict semantic concepts and integrate them into the report generation process. To learn the relationship between medical images and medical terminologies KdTNet [52] was proposed which uses a visual grid and convolution graph to extract fine-grained visual features with a transformer-based decoder to generate semantic features.

Handling data is one of the most common issues in training vision-based transformer models. The available CXR datasets are insufficient for training such heavy models. This data scarcity challenge is common across medical imaging domains [53], affecting model performance and reliability. As a result, most of these studies apply pre-trained vision transformers using ImageNet [54] weights, which does not significantly benefit the performance of these models in the medical domain [55].

C. CONTRASTIVE LEARNING

Contrastive learning has emerged as a strong alternative to vision transformers, where image-level visual representations can be effectively learned on raw image-text pairs [21]. The original model is trained on 400 million image-text pairs, demonstrating its substantial data requirements. This approach has been employed in several studies for medical reports retrieval [2], [23] using the MIMIC-CXR dataset, one of the most extensive publicly available CXR datasets, and has demonstrated improved performance in terms of F1 score accuracy. However, similar to vision transformers, the encoders used in these studies are pre-trained on natural image data, which may not effectively learn CXR features [55].

Recently multi-modal contrastive learning schemes incorporating augmentation have been proposed for natural images [25], [26], [27]. These models require fewer training data and achieve comparable accuracy to the original CLIP implementation. However, no such technique is employed for the medical report generation task.

III. PROPOSED METHOD

This study proposes the novel architecture DuCo-Net for the efficient retrieval of medical reports using CXR images through the application of contrastive learning. The core innovation lies in the synergistic combination of cross model data augmentation and joint training of contrastive learning methodologies, resulting in a dual-network scheme capable of extracting diverse features from medical images and reports. The contrastive learning used in the DuCo-Net stimulates the idea of CLIP [21]. CLIP employs a dual encoder architecture to concurrently train image and text encoders on raw image-text pairs. However, this model requires a substantial amount of data for proper training, which limits its application to small-scale datasets, especially in medical report generation where training data is limited. To address this limitation, we deployed a similar dual encoder architecture with carefully selected components tailored for the medical domain. For the visual encoder, we chose DenseNet-121 as our backbone due to its efficient feature reuse through dense connectivity, which is particularly beneficial for medical imaging tasks where fine-grained details are crucial. Although larger models like ResNet-152 offer higher capacity, they risk overfitting on limited data. Conversely, smaller models like MobileNet might not capture the subtle pathological features in chest X-rays. For the text encoder, BioBERT is leveraged pre-trained on vast biomedical corpora, offering a rich domain-specific language understanding essential for medical report comprehension. Additionally, we introduced cross-modal augmentation strategies to enhance the model's ability to learn robust representations from limited data, thereby enhancing the alignment between visual features and medical descriptions.

IV. OVERVIEW AND PRELIMINARIES

As illustrated in Fig. 2, the proposed DuCo-Net comprises two parallel contrastive learning schemes. In the primary model, which we call the backbone network, given a raw medical image-text pair (I, T), the objective is to maximize the cosine similarity between the embeddings of I and T, while minimizing the similarity between dissimilar image-text pairs. The categorical cross-entropy (CCE) loss is used to train models that match image and text pairs, such as in an image-to-text retrieval task. The goal is to align the image embeddings and text embeddings such that for a given image, the most similar text is ranked the highest, and vice versa.

Our implementation handles positive and negative pairs implicitly through batch-wise matrix computations. For a batch of size N, we compute an $N \times N$ logits matrix where

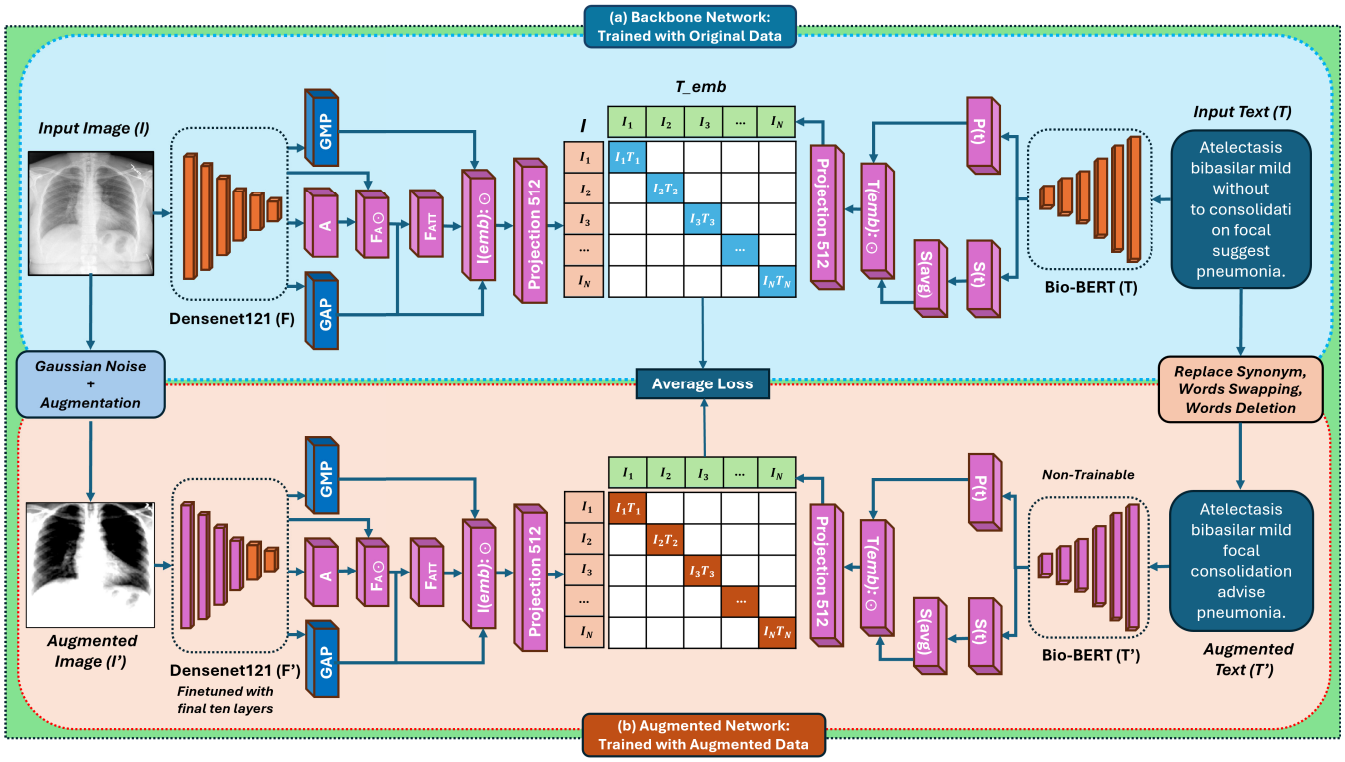


FIGURE 2. This diagram represents the detailed architecture of DuCo-Net, consisting of two main components: (a) Backbone Network, designed to betrained on original image-text pairs. (b) Augmented Network, share the same structure as the backbone network, but utilizes a combination of image andtext augmentation techniques creating multiple "views" of the original data. Additionally, enhanced pooling strategies are proposed to enrich the learningcapabilities of the Densenet121 and BioBERT encoders. Finally, the individual losses from both networks are aligned in an average loss function.

each entry (i, j) represents the scaled dot product between L2-normalized embeddings of the i-th text and j-th image. The diagonal entries (i, i) correspond to matching (positive) pairs, while off-diagonal entries represent non-matching (negative) pairs.

Formally, given a batch of N medical image-text pairs $(I_1, T_1), (I_2, T_2), \ldots, (I_N, T_N)$, we first obtain their embeddings through the backbone encoders f_I and f_T. The cross-modal logits matrix is computed as:

$$l_{ij} = \frac{f_T(T_i)^T f_I(I_j)}{\tau} \quad \forall i, j \in \{1, \ldots, N\} \tag{1}$$

where:

- $f_I(I_j)$ and $f_T(T_i)$ are L2-normalized embeddings of the j-th image and i-th text
- τ is a temperature parameter scaling the similarity scores
- The superscript T denotes vector transpose

Unlike traditional contrastive learning that uses strict binary targets (1 for matching pairs, 0 for non-matching), our approach computes soft targets by averaging intra-modal similarities S^c_{ij} and S^i_{ij}. This formulation allows related but non-matching pairs to have intermediate similarity values. For example, if two chest X-rays show different stages of pneumonia, their visual embeddings would have high similarity (S^i_{ij}), and their corresponding text descriptions would also be semantically similar (high S^c_{ij}). Consequently, their cross-modal embeddings should be more similar compared to completely different conditions. This ability to capture nuanced relationships is crucial in medical imaging, where subtle variations can have significant diagnostic implications. To implement this, we first compute comprehensive pairwise similarities within each modality, resulting in two $N \times N$ similarity matrices:

$$S^c_{ij} = f_T(T_i)^T f_T(T_j) \quad \forall i, j \in \{1, \ldots, N\} \tag{2}$$

$$S^i_{ij} = f_I(I_i)^T f_I(I_j) \quad \forall i, j \in \{1, \ldots, N\} \tag{3}$$

where:

- S^c_{ij} represents similarity between i-th and j-th text embeddings
- S^i_{ij} represents similarity between i-th and j-th image embeddings
- When $i = j$, elements represent self-similarities (diagonal)
- When $i \neq j$, elements represent cross-sample similarities (off-diagonal)

The soft targets are then computed by averaging these intra-modal similarity matrices element-wise:

$$T_{ij} = \text{softmax}\left(\frac{S^c_{ij} + S^i_{ij}}{2\tau}\right) \quad \forall i, j \in \{1, \ldots, N\} \tag{4}$$

Using these $N \times N$ logits and targets, we compute the bidirectional cross-entropy loss for the backbone network:

$$L_{base} = -\frac{1}{2}[\sum_{i,j=1}^{N} T_{ij} \cdot \log(\frac{\exp(l_{ij}/\tau)}{\sum_k \exp(l_{ik}/\tau)}) + \sum_{i,j=1}^{N} T_{ji} \cdot \log(\frac{\exp(l_{ji}/\tau)}{\sum_k \exp(l_{ki}/\tau)})] \quad (5)$$

where:

- N is the batch size, determining the size of logits and target matrices
- The summation $\sum_{i,j=1}^{N}$ computes loss over all elements in the N $\times$ N matrices
- For each row i in the logits matrix: The diagonal element l_{ii} represents the positive pair score. The off-diagonal elements l_{ij} (where $i \neq j$) represent N-1 negative pair scores
- $\sum_k \exp(l_{ik}/\tau)$ normalizes over all possible matches for the i-th text/image
- The first term $\sum_{i,j=1}^{N} T_{ij} \cdot \log(\ldots)$ aligns texts to their matching images
- The second term $\sum_{i,j=1}^{N} T_{ji} \cdot \log(\ldots)$ aligns images to their matching texts
- The factor $\frac{1}{2}$ averages the bidirectional alignment losses

In parallel to the backbone network, we introduce a secondary contrastive learning model called the augmented network (Fig 2). This network is equipped with specialized encoders that generate augmented versions of the original image-text pairs, denoted as (I', T') for the image I and caption T. In the augmented network, rather than directly maximizing the similarity between original and augmented pairs, we measure how well the similarity of original data aligns with the targets generated from the augmented network. The key distinction is that we use the original backbone network's embeddings to compute logits and the augmented network's embeddings to compute targets. This approach helps the network learn robust and invariant feature representations by ensuring consistency between original and augmented semantic structures.

For logits, we use the backbone network's embeddings:

$$l_{ij} = \frac{f_T(T_i)^T f_I(I_j)}{\tau} \quad \forall i, j \in \{1, \ldots, N\} \quad (6)$$

For targets, we compute intra-modal similarities using augmented embeddings:

$$S_{ij}^{c'} = f'_T(T_i)^T f'_T(T_j) \quad \forall i, j \in \{1, \ldots, N\} \quad (7)$$

$$S_{ij}^{i'} = f'_I(I_i)^T f'_I(I_j) \quad \forall i, j \in \{1, \ldots, N\} \quad (8)$$

where:

- f'_I and f'_T are the specialized encoders of the augmented network
- $S_{ij}^{c'}$ and $S_{ij}^{i'}$ are similarity matrices from augmented space

The augmented targets are computed similarly to the backbone network:

$$T'_{ij} = \text{softmax}(\frac{S_{ij}^{c'} + S_{ij}^{i'}}{2\tau}) \quad \forall i, j \in \{1, \ldots, N\} \quad (9)$$

Using these $N \times N$ logits and augmented targets in a batch, the augmented network loss is defined:

$$L_{aug} = -\frac{1}{2}[\sum_{i,j=1}^{N} T'_{ij} \cdot \log(\frac{\exp(l_{ij}/\tau)}{\sum_k \exp(l_{ik}/\tau)}) + \sum_{i,j=1}^{N} T'_{ji} \cdot \log(\frac{\exp(l_{ji}/\tau)}{\sum_k \exp(l_{ki}/\tau)})] \quad (10)$$

where:

- The structure mirrors the backbone loss but uses targets T'_{ij} from augmented embeddings while keeping logits l_{ij} from original embeddings
- The summation $\sum_{i,j=1}^{N}$ again computes loss over all N $\times$ N matrix elements in a batch
- The first term uses T'_{ij} to guide how original text embeddings should align with original image embeddings, based on similarities learned in augmented space
- The second term does the same for image-to-text alignment
- The normalization $\sum_k \exp(l_{ik}/\tau)$ ensures we consider all possible matches in the original embedding space
- The factor $\frac{1}{2}$ maintains the bidirectional nature as in backbone network

This formulation encourages consistency between the similarity structures of original and augmented representations, as the augmented targets are used to guide the alignment of the original embeddings. The final loss combines both networks:

$$L_{\text{total}} = L_{base} + L_{aug} \quad (11)$$

This dual contrastive learning approach is particularly crucial in medical imaging, where variations in image quality and textual descriptions are common. Although the specialized encoders (augmented network) operate independently from the backbone network encoders, their learning signals are effectively transferred to the backbone encoders through the gradient updates driven by this combined loss L_{total}. The interaction between the two networks creates a robust learning mechanism as the backbone network learns to align image-text pairs, the augmented network ensures these alignments remain consistent under different data variations, which is critical for reliable multi-modal understanding in medical applications.

A. BACKBONE NETWORK

As illustrated in Fig.2, the backbone network consists of image and text encoders. For the image encoder, DenseNet121 is employed [30], while the text encoder utilizes BioBERT [31]. Both models are fully fine-tuned on the datasets used in this study. The features learning

capabilities of these encoders are enhanced by incorporating additional pooling layers. Specifically, DenseNet121, which is pre-trained on natural images, is adapted for (CXRs) via custom pooling methods that better capture the unique characteristics of medical imaging. The detailed architectures of both encoders are discussed in the following sections.

1) ENHANCED DENSENET121

We propose an optimized adaptation of DenseNet121 specifically tailored for CXR image feature extraction. The CXR images are more challenging to classify than natural images due to their complex anatomical structures [56]. Given an input image $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent height, width, and channels respectively, the enhanced DenseNet121 image encoder incorporates a dual-pooling strategy, combining global average pooling (GAP) and global max pooling (GMP) to capture both overall anatomical information and localized high-intensity and most prominent features crucial in medical imaging. Let $F \in \mathbb{R}^{h\times w\times d}$ be the feature maps output by the final dense block. The pooled features are computed as follows:

$$F_{GAP} = GAP(F) = \frac{1}{h\cdot w}\sum_{i=1}^{h}\sum_{j=1}^{w} F_{ij} \in \mathbb{R}^d \tag{12}$$

$$F_{GMP} = GMP(F) = \max_{i,j} F_{ij} \in \mathbb{R}^d \tag{13}$$

where F_{ij} represents the feature value at position (i, j), and d is the number of feature channels.

Further, we integrate a weighted attention mechanism $A(I)$ to focus on diagnostically relevant regions within image (I). The attention module generates a spatial attention map $A \in \mathbb{R}^{h\times w}$, and the attention-weighted features F_A are computed as:

$$A = \text{sigmoid}(\text{Conv2D}_{1\times1}(F)) \in \mathbb{R}^{h\times w\times 1} \tag{14}$$

$$F_A = A \odot F \in \mathbb{R}^{h\times w\times d} \tag{15}$$

$$F_{ATT} = GAP(F_A) \in \mathbb{R}^d \tag{16}$$

where F_{ij} represents the feature value at position (i, j), d is the number of feature channels, and $\odot$ denotes element-wise multiplication. The final image embedding I_{emb} is created by concatenating these features:

$$I_{emb} = [F_{GAP}; F_{GMP}; F_{ATT}] \in \mathbb{R}^{3d} \tag{17}$$

This comprehensive representation of I_{emb} encapsulates various aspects of the medical images including the global context, local high-intensity features, and attention-weighted information. The resulting image embedding is then projected and normalized to create the final representation used in the contrastive learning framework. By incorporating these essential enhancements, the DenseNet121 variant is better equipped to extract relevant features from CXR images, leading to improved performance in medical image-text retrieval tasks.

2) ENHANCED BIOBERT

Building upon the BioBERT model pre-trained on a large corpus of biomedical data, we introduce several changes for processing medical reports and descriptions. Given an input text T, the modified BioBERT leverages both the pooled output $P(T) \in \mathbb{R}^d$ (for overall context) and sequenced output $S(T) \in \mathbb{R}^{L\times d}$ (for detailed token-level information), where L denotes the sequence length and d indicates the hidden dimension. We applied GAP to $S(T)$, resulting:

$$S_{avg}(T) = \frac{1}{L}\sum_{i=1}^{L} S_i(T) \in \mathbb{R}^d \tag{18}$$

The purpose of the GAP is to capture an aggregate representation of token-level features to create a compact, fixed-size representation of the token-level information. This operation reduces the variable length of the sequence output $S(T)$, to a single vector, allowing it to be easily combined with the pooled output $P(T)$. By doing so, the model extracts features from the overall context $P(T)$ and the average of all token-level details $S_{avg}(T)$, providing a more detailed representation of the input text. This approach enables the model to handle variable-length sequence output $S(T)$ while retaining information from every token, potentially improving its performance on tasks involving medical reports and descriptions. The final text embedding T_{emb} is formed by concatenating these representations:

$$T_{emb} = [P(T); S_{avg}(T)] \in \mathbb{R}^{2d} \tag{19}$$

yielding a rich representation encapsulating the global context and specific medical terminologies. This combined embedding undergoes further projection and normalization to align with the image embedding space, facilitating effective cross-modal learning in the contrastive framework.

B. AUGMENTED NETWORK

The augmented network has a significant role in improving the robustness and adaptability of the backbone network for learning features from medical images and text. This additional network works alongside the primary encoders of the backbone network, processing augmented versions of the input data. This combined optimization encourages the backbone encoders to learn representations that are not only consistent across different types of data but also resilient to variations in the data. We employed a strategic parameter freezing approach to preserve the valuable features learned during large-scale pre-training while adapting model to the medical domain. In the DenseNet121 architecture, only the final ten layers are fine tuned, keeping the weights of earlier layers frozen. The choice of ten layers was determined through extensive experiments, including fully training the network, completely freezing it, and training with just the last ten layers. However, it remains uncertain whether this number of layers is optimal. Research indicates that large models perform better when fine-tuned on the last layers [57]. Detailed analysis can be found in Section V-E3.

The BioBERT component of the augmented network remains entirely frozen to maintain its pre-trained weights (see Fig. 2). This selective freezing strategy helps prevent overfitting on smaller medical datasets and reduces computational overhead during training. Training specifically targets the final few layers of the encoders and a custom projection head. The projection head consists of several dense layers with GELU activation, dropout for regularization, and residual connections. This projection mechanism is applied to both image and text embeddings, aligning the feature spaces of both modalities while fine-tuning them for the specific task.

1) ENHANCED AUGMENTED DENSENET121

The augmented DenseNet121 encoder enhances the robustness of the vision encoder by applying a series of transformations to the input image (I), resulting in an augmented version (I'). This process begins with Gaussian noise $G(I)$ of ratio 0.02%, where some portions of image (I) are randomly obscured, encouraging the model to learn from partial information. Subsequently, similar to [58] (I) undergoes a sequence of random transformations including horizontal flips $F(I)$, random rotations $R(I)$, random zooming $Z(I)$, and adjustments to contrast $C(I)$ and brightness $B(I)$. These augmentations create diverse variations of (I), simulating different imaging conditions and perspectives. Thus,

$$I' = B(C(Z(R(F(G(I)))))) \tag{20}$$

Similar to the backbone encoder, Image (I') is then processed through several multiple pooling strategies, including global average pooling (GAP), global max pooling (GMP), and an attention mechanism (ATT), culminating in a rich, comprehensive embedding $E(I')$ of the medical image:

$$E(I') = [F_{GAP}(I'); F_{GMP}(I'); F_{ATT}(I')] \tag{21}$$

2) ENHANCED AUGMENTED BIOBERT

The augmented BioBERT encoder enhances the robustness of medical reports by implementing a sophisticated augmentation and encoding pipeline. The core of this encoder lies in the text augmentation process, applying a series of transformations to the input text (T), resulting in an augmented version (T'). This process employs multiple augmentation techniques, each applied with a 20% ratio, including synonym replacement $S(T)$, random word swapping $W_s(T)$, and random word deletion $W_d(T)$. The augmented text is then padded to a fixed length (L) to ensure consistency:

$$T' = \mathrm{P}_L(\mathrm{Tr}_L(W_d(W_s(S(T))))) \tag{22}$$

where P_L denotes the padding to length (L), and Tr_L denotes the truncation to length (L). The augmented text (T') is then processed via the BioBERT model, generating pooled and sequence outputs. After applying GAP to the sequenced output, these outputs are combined through concatenation, creating a rich, contextual representation. Finally, this combined representation undergoes projection through multiple dense layers with dropout, resulting in the final text embedding $E(T')$:

$$E(T') = \Phi([P(T'); Savg(T')]) \tag{23}$$

where Φ denotes the projection function, $P(T')$ reperesents the pooled output, and $Savg(T')$ indicates the GAP of the sequenced output. This multi-faceted approach captures diverse linguistic and domain-specific variations, enhancing the ability of the model to understand and represent medical text data effectively.

V. EXPERIMENTAL RESULTS

A. DATASETS

This study uses two benchmark datasets, MIMIC-CXR [2] and IU-Xray [32] for training, evaluation, and testing.

1) IU-XRAY

The IU-Xray is publicly available and contains 3,955 de-identified radiology reports and impressions, each associated with frontal and lateral CXR images. The dataset includes 7,470 CXR images. This work partitioned the data into a 7:2:1 ratio for training, validation, and test sets. All evaluations are conducted on the testing set.

2) MIMIC-CXR

This MIMIC-CXR dataset consists of free-text radiology reports, including 377,110 CXR images and 227,943 reports from 225,000 studies conducted at the Beth Israel Deaconess Medical Center between 2011 and 2016. However, the data are not well-organized, because many studies were missing reports or impressions. Therefore, for the experiments, we only used 20,000 records where both findings and impressions were available. The data were partitioned for training, validation, and testing following the same ratio as IU-Xray, with a split of 7:2:1. All evaluations are performed on the testing set.

B. EXPERIMENT SETUP

1) IMPLEMENTATION

The DuCo-Net architecture is implemented with carefully chosen parameters and optimization strategies. For the imaging modality, it utilizes DenseNet121 [30], which is equipped with a 512-dimensional projection layer, comprising sequential dense layers with GELU activation, dropout, layer normalization, and residual connections. Similarly, the text modality employs BioBERT [31] with a matching 512-dimensional projection layer to ensure alignment of image and text embeddings in a shared space. Both networks share consistent architectural choices in their projection layers, featuring a dropout rate of 0.3% to prevent overfitting, a temperature parameter = 0.07 for embedding normalization, and an Adam optimizer [62] with an initial learning rate of 0.001.

TABLE 1. Showing various models and their corresponding evaluation metrics on the IU-Xray dataset. All the results are cited from [14].

Dataset	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	ROUGE-L	METEOR	F1
IU-Xray	Show-Tell [33]	0.24	0.13	0.10	0.07	0.30	-	0.15	-
	Att2in [59]	0.24	0.13	0.11	0.09	0.30	-	0.16	-
	AdaAtt [60]	0.28	0.20	0.15	0.12	0.31	-	0.16	-
	M^2Trans [61]	0.40	0.28	0.16	0.14	0.32	-	0.17	-
	R2Gen [12]	0.47	0.30	0.21	0.16	0.37	-	0.18	-
	METransformer [14]	0.48	0.32	0.22	**0.17**	0.38	-	0.19	-
	DuCo-Net (*Ours*)	**0.50**	**0.33**	**0.22**	0.16	**0.40**	**0.26**	**0.24**	**0.40**

TABLE 2. Showing various models and their corresponding evaluation metrics on the MIMIC-CXR dataset. Results with * indicate the retrieval methods and are directly cited from their respective papers, while for the other papers, results are taken from [14].

Dataset	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	ROUGE-L	METEOR	F1
MIMIC-CXR	CXR-RePaiR-2* [22]	-	0.06	-	-	-	-	-	0.25
	CXR-RePaiR-Select* [22]	-	0.05	-	-	-	-	-	0.27
	M^2Trans [61]	0.33	0.21	0.14	0.10	0.26	-	0.13	-
	CXR-IRGEN(F)* [23]	0.32	0.16	0.10	0.06	-	0.20	-	0.29
	R2Gen [12]	0.35	0.21	0.14	0.10	0.27	-	0.14	-
	METransformer [14]	0.38	0.25	0.16	0.12	0.29	0.15	-	-
	DuCo-Net (*Ours*)	**0.42**	**0.27**	**0.17**	**0.12**	**0.34**	**0.22**	**0.20**	**0.34**

The training process involves different parameter-updating strategies across the networks. In the backbone network, both DenseNet121 and BioBERT are fully fine-tuned while the projection layers are trained from scratch. Conversely, in the augmented network, the trainable parameters of DenseNet121 are limited to the final ten layers, whereas BioBERT remains completely frozen. This freezing strategy for BioBERT is particularly effective as it leverages the model's robust pre-training on extensive biomedical corpora, preserving its ability to understand complex medical terminologies and semantic relationships while preventing catastrophic forgetting of limited domain-specific data. The projection layers maintain separate parameters from the backbone network, ensuring no weight sharing between the two networks. Although the backbone and augmented networks maintain separate parameters without weight sharing, they are unified through a combined loss function. The loss from the augmented network influences the training of the backbone network through gradient updates, despite the frozen state of most other parameters in the augmented network.

For the IU-Xray dataset, training was conducted with a batch size of 100 on 7,470 samples, where 70% (5,229 samples) were used for training and the remaining 30% for validation and testing. The model is trained for a total number of 50 epochs resulting in approximately 53 iterations per epoch (2,650 iterations total). For the larger MIMIC-CXR dataset, we utilized 20k image-text pairs with the same batch size and epoch settings, maintaining the same 70-30 split ratio for training, validation, and testing, which resulted in approximately 140 iterations per epoch (7,000 iterations total). This implementation strategy ensures efficient feature learning while preventing overfitting on both datasets, with the dual-network approach and differential parameter updating providing a comprehensive learning framework that captures both general and domain-specific features in medical imaging and report generation tasks.

2) EVALUATION METRICS

Following the standard evaluation protocol, we applied the most widely used metrics for text evaluation including BLEU [63], METEOR [64], ROUGE-L [65], and F1 to evaluate the quality of the retrieved reports on the testing dataset. These multiple complementary evaluation metrics comprehensively assess the report quality and clinical accuracy of our model. The BLEU metric evaluates n-gram precision by comparing word sequences between the retrieved and reference reports, with BLEU-1 through BLEU-4 examining increasingly longer phrase matches (1-4 words). This progression helps assess both vocabulary accuracy and proper phrase construction in medical reporting. The ROUGE-L metric focuses on recall by measuring the longest common subsequence between reports, effectively capturing how well the retrieved report maintains the sequential structure and completeness of the reference report. This is particularly important for preserving the logical flow of medical observations. METEOR extends beyond exact matches by recognizing synonyms and paraphrases, addressing the inherent variability in how radiologists may describe identical findings. For instance, terms like opacity

and consolidation might be used interchangeably in certain contexts. The F1 score derived from CheXpert [66] labeling provides a clinical effectiveness measure by evaluating the accuracy of abnormality detection in the retrieved reports against the reference reports. This metric specifically assesses whether clinically significant findings are accurately preserved, making it a crucial indicator of the model's practical utility in medical settings.

C. RESULTS

We completed a thorough assessment of the proposed model in comparison to the leading medical report generation and retrieval systems. The proposed method performed better than all baseline models across two key metrics: traditional NLG metrics and the F1 score. The F1 score was calculated from labels generated by the CheXpert [66] labeler for the original and retrieved reports. The detailed results of the evaluation are presented in Tables 1 and 2, based on the testing data, reflecting high accuracy. For the image-text retrieval component, we used pre-trained encoders from the backbone network to create embeddings for images and reports in the testing sets. We then computed the cosine similarity between the embeddings of test images and training reports to identify the most relevant reports, which were subsequently compared with the testing data reports to evaluate performance. While IU-Xray is a relatively smaller dataset with 7,470 CXR images, it provides well-structured reports and consistent imaging protocols, enabling DuCo-Net to achieve higher performance metrics (BLEU-1: 0.50, ROUGE: 0.40, F1: 0.40). In contrast, MIMIC-CXR represents a more challenging and diverse dataset with 20,000 selected records, where performance metrics show expected moderation (BLEU-1: 0.42, ROUGE: 0.34, F1: 0.34) due to greater variability in reporting styles and imaging conditions.

FIGURE 3. Bootstrap analysis (n = 1000) of model performance metrics across IU-Xray and MIMIC-CXR datasets, showing 95% confidence intervals and statistical significance ($p < 0.001$).

Despite this natural performance variation between datasets, DuCo-Net consistently outperforms existing approaches on both datasets, demonstrating its robustness across different clinical settings. The performance difference between these datasets highlights how institutional factors such as report structure, terminology variation, and imaging protocols can influence results, while also showcasing the model's ability to maintain competitive performance even in more diverse and challenging real-world clinical settings.

D. STATISTICAL VALIDATION AND PERFORMANCE ANALYSIS

To establish the reliability of our results, we conducted extensive bootstrap analysis with 1000 iterations across both IU-Xray and MIMIC-CXR datasets as shown in (Fig. 3). Our model demonstrated statistically significant improvements ($p < 0.001$) across all metrics in both datasets. For the IU-Xray dataset, we observed robust performance with narrow confidence intervals: BLEU (0.50 [0.483-0.517]), ROUGE (0.40 [0.384-0.416]), METEOR (0.24 [0.226-0.254]), and F1 (0.40 [0.385-0.415]). The MIMIC-CXR dataset similarly showed strong performance: BLEU (0.42 [0.404-0.436]), ROUGE (0.34 [0.326-0.354]), METEOR (0.20 [0.187-0.213]), and F1 (0.34 [0.326-0.354]). The consistently narrow confidence intervals across both datasets (typically ±1.5-2.0%) indicate stable model performance regardless of the data source, while the maintained statistical significance across different medical imaging contexts reinforces the model's generalizability. The systematic performance pattern, with IU-Xray showing approximately 15-20% higher scores across metrics, reflects dataset-specific characteristics while maintaining robust relative improvements over baseline approaches.

E. ABLATION STUDY

This study on medical report retrieval employs a multifaceted approach. The key contributions include the proposed modified encoders and unified training framework to exploit intrinsic data properties in each modality and extract meaningful semantic information from cross-modal correlation. For this purpose, we conducted an ablation study to determine the contributions of each component in DuCo-Net in detail.

TABLE 3. Performance comparison of various pooling strategies in Densenet121 using the IU-Xray dataset.

GAP	GMP	GAP ⊕ GMP	Attention	F1 Score
✓				0.34
	✓			0.37
		✓		0.38
✓	✓	✓	✓	**0.40**

1) EVALUATION OF THE MODIFICATIONS IN THE PROPOSED DENSENET121

We evaluated DenseNet121 with various combinations of the proposed components to assess their individual and combined contributions. We conducted experiments by systematically adding these components and measuring the performance using the F1 score. Results as shown in

TABLE 4. Table showing various Components of DuCo-Net with corresponding evaluation metrics on the IU-Xray dataset.

Dataset	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	ROUGE-L	METEOR	F1 Score
IU-Xray	Backbone network	0.42	0.26	0.16	0.09	0.30	0.19	0.18	0.32
	Augmented network	0.44	0.28	0.18	0.11	0.34	0.20	0.19	0.34
	DuCo-Net	**0.50**	**0.33**	**0.22**	**0.16**	**0.40**	**0.26**	**0.24**	**0.40**

Table. 3, indicates that the highest F1 score was achieved using DenseNet121 with all components: GAP, GMP, the concatenation, and attention mechanism. The baseline model with only GAP achieved an F1 score of 0.34. Similarly using GMP only improved the F1 score to 0.37, suggesting that the combination of GMP captures more comprehensive features from CXR images. Combining GAP and GMP using concatenation achieved an F1 score of 0.38. Further, incorporating the attention mechanism boosted the F1 score to 0.40, indicating that focusing on relevant regions enhances the discriminative power of the model. The result demonstrates that combining diverse pooling strategies and attention allows the model to capture global patterns and local details while emphasizing the most relevant areas in CXR images. These experiments were recorded while training DuCo-Net.

2) EVALUATION OF PERFORMANCE COMPARISON OF DUCO-NET AND COUNTERPARTS

The ablation study for the IU-Xray dataset is detailed in Table 4, illuminating the efficacy of the architecture of DuCo-Net and its constituent elements. The experimental approach involved training the backbone and augmented networks separately by fine-tuning them using all layers. Afterward, we integrated these networks using a joint training framework with the proposed freezing strategy for the augmented network, which we've termed DuCo-Net. The modified encoders in the backbone network exhibited moderate performance (BLEU-1: 0.42 and ROUGE: 0.30). The augmented network showcased notable improvement, particularly in BLEU scores (BLEU-1: 0.44) and ROUGE (0.34), signifying the enhanced capture of textual similarity.

However, the complete DuCo-Net, amalgamating both networks in a multi-modal contrastive learning scheme, significantly surpassed the performance of the individual components across all metrics. The complete model attained the highest scores in BLEU-1 (0.50), ROUGE (0.40), METEOR (0.24), and the F1 score (0.40). To further validate the performance of the proposed DuCo-Net versus individual training networks, we conducted qualitative analysis on some of the reports retrieved by these networks compared to the ground truth reports (Table 5). The detailed results demonstrate the ability of DuCo-Net to capture specific aspects of the retrieved reports. This overall improvement across evaluation metrics highlights the synergistic influence of integrating both networks in the dual-contrastive learning framework. These results underscore the superior capability of DuCo-Net to capture intricate relationships between medical images and reports, underscoring the advantages of uniting the backbone and augmented networks in this architecture.

3) EVALUATION OF THE PARAMETER FREEZING EFFECT ON DUCO-NET PERFORMANCE

This section evaluates the performance of DuCo-Net by examining different parameter freezing strategies for the augmented network encoders. Table 6 presents the experimental results from three distinct configurations tested on the IU-Xray dataset, each utilizing varying proportions of frozen versus trainable parameters. Initial experiments revealed that allowing all layers of the Bio-BERT model (with 110,532,864 parameters) to be fine-tuned led to overfitting. The extensive number of trainable parameters in Bio-BERT can cause the model to focus too narrowly on limited training data instead of learning generalizable features. Consequently, the parameters of the Bio-BERT model were kept frozen across all configurations. The investigation then aimed to optimize the training configuration of DenseNet121, which consists of 7,995,940 parameters. In the first configuration, DenseNet121 is fine tuned with all parameters while keeping Bio-BERT model completely frozen, resulting in BLEU-1, ROUGE, and F1 scores of 0.45, 0.38, and 0.36 respectively. These scores indicate that allowing too much flexibility in the vision encoder can lead to suboptimal feature extraction, as the model may overfit to specific visual patterns in the training data. The second configuration, with parameters of both encoders entirely frozen, showed improved performance with scores of 0.47, 0.39, and 0.37. This improvement indicates that the pre-trained weights of DenseNet121 effectively capture useful visual features, and preventing any modification helps maintain these robust features. However, the completely frozen state limits the model's ability to adapt to dataset-specific characteristics. The third configuration achieved the best performance by adopting a hybrid approach: maintaining Bio-BERT in a frozen state while selectively unfreezing 5,020,130 parameters in DenseNet121's final layers, keeping the remaining 7,493,927 parameters frozen. This strategic parameter configuration yielded the highest scores across all metrics (BLEU-1: 0.50, ROUGE: 0.40, F1: 0.40). This performance can be attributed to two factors: (1) the frozen early layers preserve essential low-level visual features from pre-training, (while (2) the later) trainable layers allow the model to adapt its high-level

feature extraction to the specific patterns in medical images. These results demonstrate that selective parameter unfreezing in the vision encoder, combined with a completely frozen language model, provides the optimal configuration for the DuCo-Net architecture.

4) COMPARATIVE ANALYSIS OF VARIOUS BATCH AND PROJECTION SIZES

We conducted an analysis to examine how batch size and the dimensionality of the projection layer affect the performance of the proposed model using the IU-Xray dataset. Various batch sizes (32, 64, and 100) and projection layer dimensionalities (256, 512, and 1024) are explored during experiments. The results, summarized in Fig. 4, indicate a consistent improvement in performance with larger batch sizes and higher projection dimensionalities. Specifically, the F1 scores for the findings (reports) task increased significantly from 32 (using a batch size of 32 and a projection dimension of 256) to 40 (with a batch size of 100 and a projection dimension of 512). This indicates that larger batch sizes yield more stable gradient estimates, while higher-dimensional projections capture more detailed features from the medical data. However, we observed a slight degradation in performance when increasing the projection dimension to 1024. This indicates that contrastive learning may struggle to capture rich features when the projection size is excessively large. These findings highlight the importance of carefully tuning these hyper-parameters to optimize model performance in medical image analysis tasks.

VI. DISCUSSION

The results and detailed ablation study demonstrate the effectiveness of our proposed DuCo-Net model. The robustness to real-world variability in medical imaging data is enhanced through comprehensive data augmentation strategies during training. For the imaging modality, we applied various augmentation techniques to simulate real-world conditions commonly encountered in clinical settings. These include contrast adjustments to account for varying exposure levels, random rotations to handle patient positioning variations, brightness modifications, geometric transformations like scaling and translation and Gaussian noise injection. These augmentations help the model learn invariant features that are robust to common image quality variations in clinical practice. For the text modality, we leveraged BioBERT's domain-specific knowledge and incorporated text augmentation to ensure our model remains robust to variations in real-world data. The effectiveness of this comprehensive strategy is evidenced by our experimental results, where the model maintained consistent performance despite the introduced variations in both modalities. This demonstrates the model's resilience to both image quality variations and textual variations that might be encountered in real-world clinical settings. Our optimization studies with different batch sizes and projection

TABLE 5. Comparison of report retrieval performance between DuCo-Net and Counterpart networks utilizing the IU-Xray [32] dataset as the ground truth. The report sentences are highlighted in distinct colors to enhance visualization. Sentences that are highlighted in the same color for each model when compared to the ground truth report indicate accurate retrieval of information.

Image	Ground Truth	Backbone-Network	Augmented-Network	DuCo-Net
	The heart is normal in size. The mediastinum is unremarkable. XXXX XXXX opacities in right mid lung. The lungs are otherwise grossly clear.	Borderline heart size. Elevated left diaphragm. Clear right lung. Tracheostomy tube tip above the carina Extensive airspace disease in the left base. No large effusion or pneumothorax.	The heart is normal in size and contour. The lungs are clear, without evidence of infiltrate. There is no pneumothorax or effusion.	The heart is normal in size. The mediastinum is unremarkable. The lungs are grossly clear. Underlying emphysematous changes are noted.
	The heart size and cardiomediastinal silhouette are normal. There is hyperexpansion of the lungs with flattening of the hemidiaphragm. There is no focal airspace opacity, pleural effusion, or pneumothorax. There multilevel degenerative changes of thoracic spine.	The heart and lungs have XXXX XXXX in the interval. Both lungs are clear and expanded. Heart and mediastinum normal.	The trachea is midline. The cardiomediastinal silhouette is normal. The lungs are clear, without evidence of focal infiltrate or effusion. There is no pneumothorax. The visualized bony structures reveal no acute abnormalities.	The heart size and cardiomediastinal silhouette are normal. The lungs are clear without focal airspace opacity, pleural effusion, or pneumothorax. There are numerous calcified granuloma in the right perihilar region. There are multilevel degenerative changes in the thoracic spine.
	Lungs are overall hyperexpanded with flattening of the diaphragms. No focal consolidation. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour. Degenerative changes in the thoracic spine.	The heart and lungs have XXXX XXXX in the interval. Both lungs are clear and expanded. Heart and mediastinum normal.	The heart, pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia.	Lungs are mildly hyperexpanded. The lungs are clear. There is no focal airspace consolidation. No pleural effusion or pneumothorax. Heart size and mediastinal contour are within normal limits. There are diffuse degenerative changes of the spine.

TABLE 6. Table summarizing various experiments conducted on the parameter freezing of DenseNet121 within the augmented network, along with their corresponding evaluation metrics on the IU-Xray dataset.

Model	Encoders	Parameters	Trainable	Non-Trainable	BLEU-1	Rouge	F1
DuCo-Net	DenseNet121	7,995,940	7,995,940	-	0.45	0.38	0.36
	BioBERT	110,532,864	-	110,532,864			
	DenseNet121	7,995,940	-	7,995,940	0.47	0.39	0.37
	BioBERT	110,532,864	-	110,532,864			
	DenseNet121	7,995,940	5,020,13	7,493,927	0.50	0.40	0.40
	BioBERT	110,532,864	-	110,532,864			

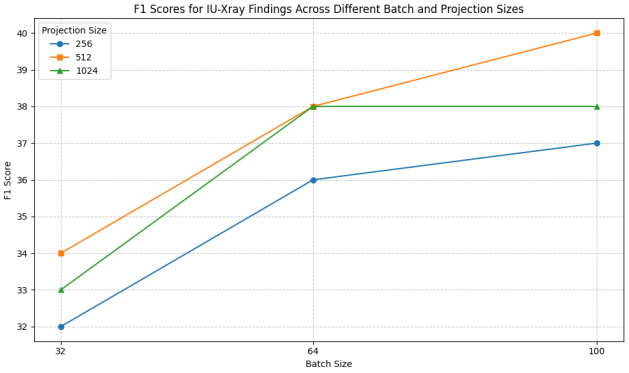


FIGURE 4. Summary of the F1 score across various batch and projection sizes.

dimensions further validate the model's robustness to input variability.

However, a fundamental limitation of DuCo-Net lies in its retrieval-based architecture. Since the model operates by retrieving and matching existing reports rather than generating new ones, its performance is intrinsically bounded by the scope and quality of the training dataset. This reliance on pre-existing reports presents challenges when encountering rare pathological conditions or unusual presentations that may be underrepresented in the training data. Additionally, the model's ability to capture subtle clinical variations is constrained by the granularity and diversity of the available training examples. While our dual-contrastive learning approach enhances feature extraction and matching capabilities, it cannot overcome the inherent limitations of retrieval-based systems in handling cases that significantly deviate from the training distribution. This limitation particularly affects the model's utility in specialized clinical scenarios or rare disease cases where appropriate reference reports might be scarce in the training corpus.

VII. CONCLUSION

This study introduces DuCo-Net, a dual-contrastive learning network designed for the efficient retrieval of medical reports using small-scale datasets. The DuCo-Net method comprises two contrastive learning models: backbone and augmented networks. The backbone model processes the original data, whereas the augmented network processes the augmented version. Both models converge on the original data using an average loss function in a unified framework. We employed a selective freezing strategy in the augmented network encoders to prevent overfitting during training on small-scale data, reducing the computational overhead. Additionally, we proposed modified versions of the DenseNet121 and BioBERT models as encoders in the DuCo-Net architecture, designed to process medical images and text reports. In addition, DuCo-Net outperforms state-of-the-art models, demonstrating superior performance on standard NLG metrics and F1 scores. This finding indicates its capability to generate accurate and relevant medical reports. The success of the proposed model has significant implications for automating and expediting the radiological reporting process, potentially reducing the workload for radiologists and improving patient care.

In the future, this work could be extended to incorporate retrieval-augmented generation. Additionally, future research could explore applying DuCo-Net to other medical imaging modalities, such as anomaly classification, and investigate its potential in healthcare-related NLP tasks, further advancing the field of artificial intelligence-assisted medical diagnostics.

REFERENCES

[1] *Communicating Radiation Risks in Paediatric Imaging: Information to Support Health Care Discussions About Benefit and Risk*, World Health Org., Geneva, Switzerland, 2016.

[2] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[3] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, and D. Capurro, "A survey on deep learning and explainability for automatic report generation from medical images," *ACM Comput. Surv.*, vol. 54, no. 10, pp. 1–40, Jan. 2022.

[4] P. Kisilev, E. Walach, E. Barkan, B. Ophir, S. Alpert, and S. Y. Hashoul, "From medical image to automatic medical report generation," *IBM J. Res. Develop.*, vol. 59, nos. 2–3, pp. 1–2, Mar. 2015.

[5] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Granada, Spain. Cham, Switzerland: Springer, Jan. 2018, pp. 457–466.

[6] S. Li, W. Wang, J. Li, and J. Lin, "Study on medical image report generation based on improved encoding–decoding method," in *Proc. 15th Int. Conf. Intell. Comput.*, Nanchang, China. Cham, Switzerland: Springer, Jan. 2019, pp. 686–696.

[7] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 728–737.

[8] G. O. Gajbhiye, A. V. Nandedkar, and I. Faye, "Automatic report generation for chest X-ray images: A multilevel multi-attention approach," in *Proc. 4th Int. Conf. Comput. Vis. Image Process.*, Jaipur, India. Cham, Switzerland: Springer, Jan. 2020, pp. 174–182.

[9] V. Tiwari, K. Bapat, K. R. Shrimali, S. K. Singh, B. Tiwari, S. Jain, and H. K. Sharma, "Automatic generation of chest X-ray medical imaging reports using LSTM-CNN," in *Proc. Int. Conf. Data Sci., Mach. Learn. Artif. Intell.*, Aug. 2021, pp. 80–85.

[10] E. Pahwa, D. Mehta, S. Kapadia, D. Jain, and A. Luthra, "MedSkip: Medical report generation using skip connections and integrated attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3409–3415.

[11] F. F. Alqahtani, M. M. Mohsan, K. Alshamrani, J. Zeb, S. Alhamami, and D. Alqarni, "CNX-b2: A novel CNN-transformer approach for chest X-ray medical report generation," *IEEE Access*, vol. 12, pp. 26626–26635, 2024.

[12] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," 2020, *arXiv:2010.16056*.

[13] Y. Miura, Y. Zhang, E. B. Tsai, C. P. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," 2020, *arXiv:2010.10042*.

[14] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11558–11567.

[15] D. Gao, M. Kong, Y. Zhao, J. Huang, Z. Huang, K. Kuang, F. Wu, and Q. Zhu, "Simulating doctors' thinking logic for chest X-ray report generation via transformer-based semantic query learning," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 102982.

[16] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," *Artif. Intell. Med.*, vol. 144, Oct. 2023, Art. no. 102633.

[17] W. Boag, T.-M. H. Hsu, M. B. A. McDermott, G. Berner, E. Alesentzer, and P. Szolovits, "Baselines for chest X-ray report generation," in *Proc. Mach. Learn. Health Workshop*, Jan. 2019, pp. 126–140.

[18] W. Liang, Y. Yuan, H. Ding, X. Luo, W. Lin, D. Jia, Z. Zhang, C. Zhang, and H. Hu, "Expediting large-scale vision transformer for dense prediction without fine-tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 35462–35477.

[19] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 103000.

[20] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.

[22] M. Endo, R. Krishnan, V. Krishna, A. Y. Ng, and P. Rajpurkar, "Retrieval-based chest X-ray report generation using a pre-trained contrastive language-image model," in *Proc. Mach. Learn. Health*, 2021, pp. 209–219.

[23] J. Shentu and N. A. Moubayed, "CXR-IRGen: An integrated vision and language model for the generation of clinically accurate chest X-ray image-report pairs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5212–5221.

[24] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari, "RoentGen: Vision-language foundation model for chest X-ray generation," 2022, *arXiv:2211.12737*.

[25] N. Mu, A. M. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2021, pp. 529–544.

[26] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6995–7004.

[27] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 4904–4916.

[28] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," 2021, *arXiv:2110.05208*.

[29] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, and B. Piot, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[31] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[32] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, May 2012, pp. 84–90.

[35] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.

[36] D. Lyndon, A. Kumar, and J. Kim, "Neural captioning for the ImageCLEF 2017 medical image challenges," in *Proc. CLEF*, Jan. 2017, pp. 1–15.

[37] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. Chapman, T. Amrhein, D. Mong, D. L. Rubin, O. Farri, and M. P. Lungren, "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," *Artif. Intell. Med.*, vol. 97, pp. 79–88, Jun. 2019.

[38] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, L. Rubini, M. Ruiz, G. Singh, F. Stemmer, W. Sun, P. Vozila, T. Lin, and R. Ramamurthy, "Generating medical reports from patient-doctor conversations using sequence-to-sequence models," in *Proc. 1st Workshop Natural Lang. Process. Med. Conversations*, 2020, pp. 22–30.

[39] M. Y. Ansari, Y. Yang, S. Balakrishnan, J. Abinahed, A. Al-Ansari, M. Warfa, O. Almokdad, A. Barah, A. Omer, A. V. Singh, P. K. Meher, J. Bhadra, O. Halabi, M. F. Azampour, N. Navab, T. Wendler, and S. P. Dakua, "A lightweight neural network with multiscale feature enhancement for liver CT segmentation," *Sci. Rep.*, vol. 12, no. 1, p. 14153, Aug. 2022.

[40] M. Y. Ansari, Y. Yang, P. K. Meher, and S. P. Dakua, "Dense-PSP-UNet: A neural network for fast inference liver ultrasound segmentation," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106478.

[41] M. Y. Ansari, I. A. C. Mangalote, P. K. Meher, O. Aboumarzouk, A. Al-Ansari, O. Halabi, and S. P. Dakua, "Advancements in deep learning for B-mode ultrasound segmentation: A comprehensive review," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 3, pp. 2126–2149, Jun. 2024.

[42] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998.

[43] X. Zeng, L. Wen, B. Liu, and X. Qi, "Deep learning for ultrasound image caption generation based on object detection," *Neurocomputing*, vol. 392, pp. 132–141, Jun. 2020.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[45] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informat. Med. Unlocked*, vol. 24, Apr. 2021, Art. no. 100557.

[46] A. B. Amjoud and M. Amrouch, "Automatic generation of chest X-ray reports using a transformer-based deep learning model," in *Proc. 5th Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2021, pp. 1–5.

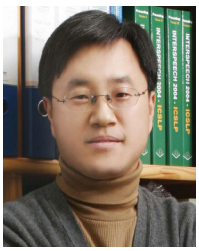
[47] G. Liu, Y. Liao, F. Wang, B. Zhang, L. Zhang, X. Liang, X. Wan, S. Li, Z. Li, S. Zhang, and S. Cui, "Medical-VLBERT: Medical visual language BERT for COVID-19 CT report generation with alternate learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3786–3797, Sep. 2021.
[48] M. Kim, K.-R. Moon, and B.-D. Lee, "Unsupervised anomaly detection for posteroanterior chest X-rays using multiresolution patch-based self-supervised learning," *Sci. Rep.*, vol. 13, no. 1, p. 3415, Feb. 2023.
[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[50] D. You, F. Liu, S. Ge, X. Xiao-Xia, J. Zhang, and X. Wu, "AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Strasbourg, France. Cham, Switzerland: Springer, Jan. 2021, pp. 72–82.
[51] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2022, pp. 655–664.
[52] Y. Cao, L. Cui, F. Yu, L. Zhang, Z. Li, N. Liu, and Y. Xu, "KdTNet: Medical image report generation via knowledge-driven transformer," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, Jan. 2022, pp. 117–132.
[53] M. Y. Ansari, M. Qaraqe, R. Righetti, E. Serpedin, and K. Qaraqe, "Unveiling the future of breast cancer assessment: A critical review on generative adversarial networks in elastography ultrasound," *Frontiers Oncol.*, vol. 13, Dec. 2023, Art. no. 1282536.
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[55] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 1–8.
[56] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest X-ray analysis: A survey," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102125.
[57] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–16.
[58] M. M. A. Monshi, J. Poon, V. Chung, and F. M. Monshi, "CovidXrayNet: Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104375.
[59] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.
[60] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.
[61] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[63] K. Papineni, S. Roukos, T. Ward, and W. B. Zhu, "A method for automatic evaluation of machine translation," in *Proc. ACL*, Philadelphia, PA, USA, 2001, pp. 1–11.
[64] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Jun. 2005, pp. 65–72.
[65] C.-Y. Lin and E. Hovy, "Manual and automatic evaluation of summaries," in *Proc. ACL Workshop Autom. Summarization*, vol. 4, 2002, pp. 45–51.
[66] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, and K. Shpanskaya, "A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 1–12.

ZAHID UR RAHMAN received the M.S. degree in computer science from COMSATS University, Islamabad, Pakistan, in 2023. He is currently pursuing the Ph.D. degree with the Department of Intelligent Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include machine learning, deep learning, computer vision, and medical imaging.

JU-HWAN LEE received the B.S. degree from the Department of Earth and Environmental Sciences, Chonnam National University, South Korea. He is currently pursuing the integrated Ph.D. degree with the Department of Intelligent Electronics and Computer Engineering, Chonnam National University. His research interests include deep learning, computer vision, and knowledge distillation.

DANG THANH VU received the B.S. degree in mathematics and computer science from Ho Chi Minh University of Science, Vietnam, in 2019, and the Ph.D. degree in ICT convergence engineering systems from Chonnam National University, South Korea, in 2024. He is currently an Artificial Intelligence Researcher with AISeed Inc., South Korea. His research interests include deep learning, computer vision, and capsule networks.

IQBAL MURTZA received the B.Sc. degree in mathematics and physics from GC University Faisalabad, Pakistan, in 2006, the M.Sc. and M.Phil. degrees in electronics from the Department of Electronics, Quaid-i-Azam University, Islamabad, Pakistan, in 2006 and 2011, respectively, and the Ph.D. degree in computer science from the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad, in 2018. He has been an Assistant Professor with the Faculty of Computing and Artificial Intelligence, Air University Islamabad, since May 2018. He is currently a Postdoctoral Researcher with Chonnam National University, Gwangju, South Korea. His areas of expertise are machine learning, brain-inspired modeling, mathematical and statistical modeling for data science, and digital image processing.



JIN-YOUNG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in engineering from Seoul National University, Seoul, South Korea, in 1986, 1988, and 1994, respectively. Since 1995, he has been a Professor with the Department of Intelligent Electronic and Computer Engineering, Chonnam National University, South Korea. His research interests include machine learning, signal processing, and deep learning.