KCLVA: Knowledge-enhanced Contrastive Learning and View-specific Attention for Chest X-ray Report Generation

Jinlong Zhu, and Ping Lu

University of Leeds, Leeds, UK
bfbs2497@leeds.ac.uk, p.lu@leeds.ac.uk

Abstract. In clinical scenarios, radiologists analyse multiple chest X-ray (CXR) images from various view positions to identify diseases and abnormalities. To replicate the diagnostic approach of experienced radiologists, we propose an encoder-decoder-based CXR report generation architecture, KCLVA, which leverages the Unified Medical Language System (UMLS) to extract view-specific information from diagnostic reports, focusing on posteroanterior, anteroposterior, and lateral views. This extracted information facilitates view-specific attention (VA) mechanisms and is subsequently used to construct a similarity matrix that enables many-to-many contrastive learning. In the encoder, we employ a knowledge distillation architecture to guide the learning of the student model by freezing the teacher model. Within the student text encoder, the VA mechanism is utilised to automatically assign higher weights to tokens corresponding to a specific view in diagnostic reports based on the view position of the CXR, while assigning lower weights to other tokens. The image and text features are then integrated using contrastive learning. In the decoder, a transformer-based backbone architecture is employed to decode the encoder output and generate a medical diagnosis report. This strategy leverages UMLS to extract view-specific information, employs VA to adjust token weights, and utilises many-to-many contrastive learning through a weighted contrastive loss. Together, these components enable our model to closely simulate the diagnostic process of professional radiologists. Consequently, our method achieves significant improvements of 0.185 on METEOR and 0.078 on ROUGE compared to previous approaches.

Keywords: Chest X-ray Report Generation · Contrastive Learning · Knowledge Distillation · Unified Medical Language System · View-specific Attention

1 Introduction

Chest X-ray (CXR) images are extensively utilised in medical practice, with approximately 500,000 images requested by physicians annually in the Netherlands alone [29]. While CXR images effectively reflect chest conditions, physicians require substantial expertise to accurately identify abnormalities and produce diagnostic reports. To alleviate this burden on physicians, automated radiology

report generation systems aim to generate reports directly from radiographs. Current research primarily focuses on multimodal learning approaches, including contrastive learning [13, 16, 26, 36, 40], multi-view CXR images fusion [10, 12, 37, 42, 22], and knowledge enhancement strategies [5, 17, 24, 41].

In clinical practice, as abnormalities often occupy only a small portion of each radiograph, physicians must systematically analyse multi-view CXR images—including posteroanterior (PA), anteroposterior (AP), and lateral (LTA) views—to develop effective treatment plans. Our proposed KCLVA model emulates this approach through an encoder-decoder architecture designed to learn from paired view-specific reports and radiographs.

Each sample in standard CXR datasets, such as IU-Xray [8] and MIMIC-CXR [11], typically contains one report associated with multiple CXR images from different view, each conveying distinct information. However, the report for each sample provides a comprehensive summary of multiple images, without assigning individual diagnoses to each image. To address this limitation, we extract view-specific medical terms from reports prior to training and align them with corresponding radiographs using view-specific attention (VA). These terms are then utilised to construct a similarity matrix via text similarity calculations.

Given the specialised nature of medical terminology, we incorporate the Unified Medical Language System (UMLS) [4] to enhance both the quantity and accuracy of extracted medical terms. Our model employs knowledge distillation to learn from pre-trained clinical encoders, with the VA automatically assigning higher weights to view-specific terms in reports. To address the many-to-many relationship between images and reports in CXR datasets [33], we utilise many-to-many contrastive learning for modality fusion. Our contributions include:

- We are the first to propose a novel architecture that utilises the Unified Medical Language System to extract medical terms from original reports as view-specific guided terms, leveraging these terms to construct a similarity matrix. This architecture can be applied to other datasets with multi-view X-ray images, provided that each patient is associated with a single report.
- We are the first to introduce the view-specific attention mechanism, which directs the model to assign weights to words based on extracted medical terms in diagnostic reports. This novel approach enables models to learn more effectively by emphasising view-specific terms in diagnostic reports.
- We propose a many-to-many contrastive learning objective function, weighted contrastive loss, which consists of structured matched loss and soft contrastive loss. As demonstrated by our experiments, the many-to-many contrastive learning approach enhances model performance effectively.

2 Related Work

Multi-view Chest X-ray report generation Multi-view CXR report generation has emerged as a significant research area with promising results. Zhu & Feng [42] introduced MVC-Net, which employs separate networks for different radiograph views and an additional network for feature fusion. However, their

model could not adaptively adjust to the varying importance of pathologies across views. Rubin et al. [27] trained separate CNN models for different view positions but achieved limited success. Yang et al. [37] proposed a multi-view encoder for AP and LTA X-rays that leveraged complementary information but focused solely on image-level fusion.

While these approaches have advanced multi-view CXR analysis, they primarily emphasise image-level fusion without addressing text-level alignment between images and reports. The critical connection between specific views and their corresponding textual descriptions remains largely unexplored.

Chest X-ray report generation with contrastive learning Since CLIP [26] demonstrated the effectiveness of contrastive learning for multi-modal tasks, this approach has become central to CXR report generation. Yan et al. [36] developed a weakly supervised contrastive loss that identified and weighted hard negative samples. Yang et al. [38] proposed triplet sample construction with double contrast learning across modalities. Liu et al. [16] introduced aggregate and differential attention mechanisms to extract distinguishing information by contrasting input images with normal ones.

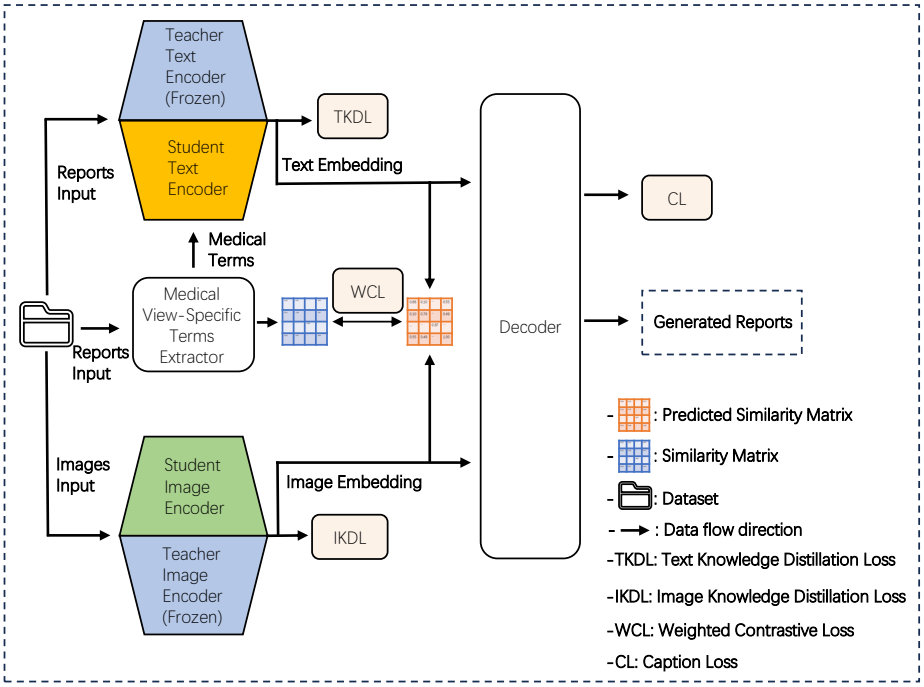
While these approaches show promise, they primarily address one-to-one or one-to-many relationships between images and text. However, they fail to capture the many-to-many relationships inherent in CXR images and reports.

Chest X-ray report generation with knowledge enhancement The specialised nature of medical diagnosis has made knowledge enhancement highly effective for CXR report generation. Liu et al. [17] integrated multiple knowledge sources and developed Case-Based and Disease-Based Retrieval mechanisms. Prabhakar et al. [24] were the first to employ UMLS for zero-shot CXR classification. Zhang et al. [41] extracted medical entities using heuristic rules, RadGraph, and ChatGPT to guide visual representation learning.

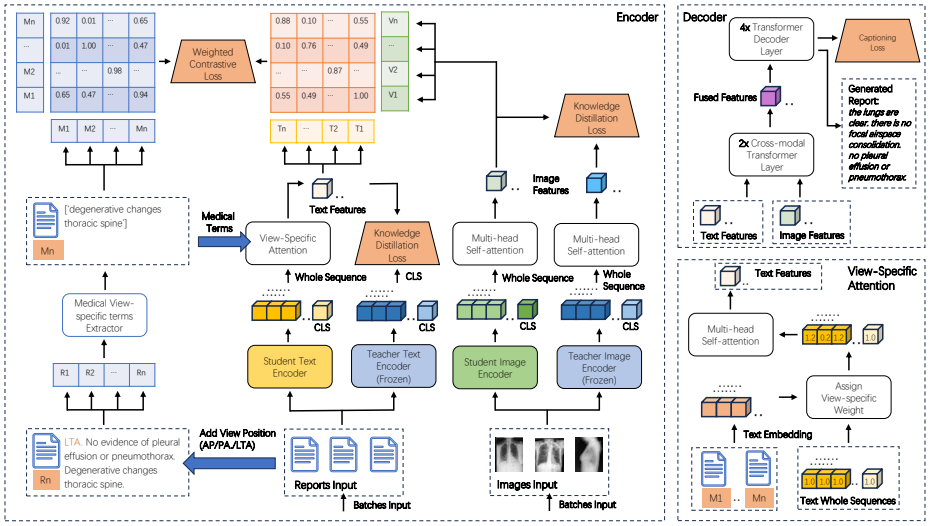
Despite these advancements, existing approaches have not fully utilised established medical knowledge bases, either relying on limited knowledge sources or applying comprehensive bases to more narrowly defined tasks than report generation.

3 Method

We present the technical details of KCLVA, following the workflow illustrated in Fig. 1. KCLVA consists of the following components: (1) a medical view-specific term extractor that identifies view-specific medical terms and constructs a similarity matrix, (2) view-specific attention, which assigns weights to words based on the extracted view-specific terms of each radiograph, (3) vision and text encoders, comprising both student and teacher encoders, (4) a decoder that integrates image and text features to generate captions for each radiograph, and (5) objective functions employed to optimise the model. The architectural details are illustrated in Fig. 1b.



(a) An overview of the proposed KCLVA workflow.



(b) An overview of KCLVA detailed architecture.

Fig. 1. The proposed KCLVA architecture consists of: (a) an overview of the KCLVA workflow and (b) the detailed KCLVA architecture. This multi-modal architecture employs a dual-encoder structure comprising frozen teacher encoders and trainable student encoders for both textual and visual pathways. The medical view-specific terms extractor identifies medical terms from input reports and constructs a similarity matrix. The architecture leverages multiple loss functions: text knowledge distillation loss (TKDL) and image knowledge distillation loss (IKDL) to transfer knowledge from teacher to student models, weighted contrastive loss (WCL) to establish many-to-many relationships between radiographs and reports, and caption loss (CL) to optimise report generation. The decoder employs a fusion layer (2x cross-modal transformer layers) to amalgamate these multi-modal representations, which are subsequently processed by 4x transformer decoder layers [30] to generate comprehensive medical reports that maintain clinical accuracy while capturing relevant visual findings.

3.1 Medical View-specific Terms Extractor

Medical View-specific Terms Extraction The proposed medical view-specific terms extractor module employs a multi-layered approach to extract view-specific medical terms from radiology reports, enabling precise alignment between CXR images and their corresponding textual descriptions. This component is essential to the KCLVA architecture, as it establishes the foundation for view-specific attention (VA) and many-to-many contrastive learning.

Our extraction methodology integrates natural language processing (NLP) with specialised medical knowledge to identify terms relevant to different radiographic views. The extraction process follows a comprehensive pipeline:

(1) Text Preprocessing: The module begins with text preprocessing to normalise the input, including case conversion, removal of redundant spaces, and correction of common abbreviations and spelling variations. This step enhances the accuracy of subsequent NLP tasks.
(2) Term Identification: A medical-domain-specific spaCy model (`en_core_sci_scibert` [19]), enhanced with custom term rules, is used to identify basic medical terms . Dependency syntax analysis is then applied to extract complete medical phrases, capturing relationships between anatomical structures and their associated findings.
(3) Pattern Matching: Regular expression pattern matching is employed to identify standardised medical expressions and complex medical phrases specific to chest radiology, such as cardiomediastinal descriptions.
(4) Negation Context Analysis: Specialised processors handle the nuanced nature of medical language by accurately identifying and representing negated findings (e.g., "no pleural effusion"). Status descriptions of anatomical structures and complex negation structures (e.g., "no X, Y, or Z") are also processed to ensure comprehensive term capture.
(5) Integration with Unified Medical Language System: A critical feature of the extractor is its integration with the UMLS knowledge base, which provides external medical knowledge for term validation and classification. UMLS enables the verification of extracted terms' medical relevance, expansion of terminology through related concepts, classification of terms based on semantic types, and view-specific filtering to identify terms particularly relevant to AP, PA, or LTA views.

Similarity Matrix Construction After extraction, post-processing steps such as term normalisation, deduplication, conflict resolution, and filtering are performed to ensure that the extracted terms accurately represent radiographic findings. The extracted view-specific medical terms are subsequently utilised to construct a similarity matrix that quantifies the relationships between different radiographs and their textual descriptions.

The extracted medical terms are mapped to their corresponding UMLS Concept Unique Identifiers (CUIs) by considering semantic types, source vocabularies, and context-aware similarity calculations to ensure precise term matching.

Using the mapped CUIs, semantic similarity between terms is computed by accounting for direct relationships such as synonyms, parent-child, and sibling relationships within the UMLS semantic type hierarchy. Furthermore, adjustments are applied to account for negation contexts and term types. The similarity scores are then used to construct a matrix where rows and columns represent terms from different CXR views. The matrix values reflect the degree of semantic similarity between each pair of terms, as represented by the following equation:

$$\text{Similarity Matrix} = \begin{bmatrix} sim(e_{11}, e_{12}) & \cdots & sim(e_{11}, e_{1n}) \\ \vdots & \ddots & \vdots \\ sim(e_{m1}, e_{12}) & \cdots & sim(e_{1m}, e_{1n}) \end{bmatrix} \tag{1}$$

where, e_{ij} represents the j-th term from the i-th CXR view, and $sim(e_{ij}, e_{kl})$ denotes the semantic similarity between term e_{ij} and term e_{kl}.

The similarity matrix is further refined by applying view-specific filtering, ensuring that only terms relevant to the specific CXR view are considered in the similarity calculations. This refinement enhances the alignment between CXR images and their associated textual descriptions.

This approach addresses a fundamental challenge in CXR datasets, where multiple images from different views are frequently associated with a single report. By extracting view-specific medical terms, fine-grained connections are established between individual radiographs and their corresponding textual descriptions, enabling more precise image-text alignment. The similarity matrix derived from these extracted terms serves as the foundation of our many-to-many contrastive learning approach, facilitating the identification of potential positive samples and the definition of appropriate margins for negative samples.

3.2 View-specific Attention

The View-specific Attention (VA) within the KCLVA architecture is designed to dynamically allocate attention weights to medical terms in radiology reports based on their relevance to specific CXR views. This mechanism is essential for improving the alignment between images and textual descriptions by focusing on view-specific information while simultaneously considering the global context of the report.

The first step involves tokenisation and position tracking. The module tokenises the reports and medical terms, tracking token positions to ensure accurate attention distribution. Let T represent the set of all tokens in the report, and S denote the set of view-specific medical terms: $T = \{t_1, t_2, \ldots, t_n\}$, $S = \{s_1, s_2, \ldots, s_m\}$, where n is the total number of tokens, and m is the number of medical terms.

The next step is the computation of attention weights. A relevance mask is calculated for each token t_i in the report, assigning higher weights to tokens within medical terms that correlate with the current CXR view. Tokens outside these terms receive lower but non-zero weights, enabling the model to capture the global context:

$$\text{Relevance Mask}(t_i) = \begin{cases} 1.0, & \text{if } t_i \in S_{\text{view-specific}} \\ \lambda, & \text{otherwise} \end{cases} \tag{2}$$

where λ is a learnable parameter representing the base weight for tokens outside the view-specific sentences.

Terms identified as medically significant (e.g., derived from UMLS) are assigned an additional weight boost to emphasise their importance:

$$\text{Medical Term Weight}(t_i) = \mu \times \text{Relevance Mask}(t_i) \tag{3}$$

where μ is a learnable parameter that enhances the weight of medical terms.

The module computes self-attention scores among tokens, adjusting these scores based on the relevance masks and medical term weights:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M\right) \tag{4}$$

where, Q, K, and V are the query, key, and value matrices derived from the input tokens, d_k is the dimension of the key vectors, and M is the matrix of attention weights derived from the relevance mask:

$$M = \text{diag}(\text{Relevance Mask}(t_1), \ldots, \text{Relevance Mask}(t_n)) \tag{5}$$

The attention weights are then applied to the value matrix:

$$\text{Attention}(Q, K, V) = AV \tag{6}$$

3.3 Vision and Text Encoder

The proposed KCLVA architecture comprises a student image encoder, a student text encoder, a teacher image encoder, and a teacher text encoder.

Vision Encoder The vision encoder in KCLVA employs a dual-architecture architecture comprising teacher and student image encoders to facilitate efficient knowledge transfer in radiographic image processing. The encoder first transforms input images $I \in R^{B\times C\times W\times H}$ through the teacher or student encoder preprocessor E_{pre} to obtain V_0. Subsequently, the images are encoded using the teacher or student vision encoder E_{img} to produce CLS_V and P_V. To leverage both global and local information, we implement an attention refinement mechanism E_{Att}, which combines the CLS token (global representation) and patch tokens (local features) to generate V_1. Finally, a projection head maps the refined embeddings into a normalised embedding space to produce V:

$$V_0 = E_{pre}(I) \tag{7}$$

$$CLS_V, P_V = E_{img}(V_0) \tag{8}$$

$$V_1 = E_{Att}(CLS_V, P_V) \tag{9}$$

$$V = f_{pv}(V_1) \tag{10}$$

where, $CLS_V \in R^{B\times 1\times d_{hidden}}$ and $P_V \in R^{B\times (N-1)\times d_{hidden}}$, while f_{pv} denotes the projection head.

Text Encoder The Text Encoder in KCLVA employs a dual-architecture architecture comprising teacher and student text encoders. It is further enhanced by a VA mechanism to optimise the representation of radiological reports. The encoder processes input text tokens $T_0 \in R^{B\times L}$ with an attention mask $M \in R^{B\times L}$ through the following sequential pipeline:

$$T_1 = E_{text}(T_0, M) \tag{11}$$

$$CLS_T, P_T = E_{VA}(T_1) \tag{12}$$

$$T = f_{pt}(P_T) \tag{13}$$

where, E_{text} represents either the teacher or student text encoder, producing hidden representations $T_1 \in R^{B\times L\times d_{hidden}}$. E_{VA} denotes the VA module, $P_T \in R^{B\times d_{hidden}}$ represents the weighted average representation, and $CLS_T \in R^{B\times d_{hidden}}$ corresponds to the CLS token representation. f_{pt} denotes the projection head.

3.4 Decoder

The Decoder in KCLVA employs a transformer-based architecture to generate radiological reports and align visual and textual representations. The decoder module processes the encoded features and generates text in an autoregressive manner:

$$L = D(V, T, T_{in}, M, \alpha_1) \tag{14}$$

where, D denotes the decoder function, T is normalised text embeddings,T_{in} represents the input token IDs, M is the attention mask, α_1 is the training phase parameter, and L corresponds to the output logits.

3.5 Objective Functions

Weighted Contrastive Loss The weighted contrastive loss (WCL) in KCLVA is designed to address the many-to-many relationships inherent in medical image-text pairs. Unlike conventional contrastive learning, which treats each image-text pair as strictly positive or negative, our approach integrates a structured matching loss and a soft contrastive loss to more effectively capture the nuanced relationships between radiological images and reports.

The structured matching loss measures the alignment between normalised prediction logits and the ground truth similarity matrix:

$$L_{str} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} w_{ij}\cdot\left(\sigma(4\cdot V_i\cdot T_j^T - S_{ij}\cdot\gamma_1 - \delta_{ij}\cdot(1-\gamma_1))\right)^2 \tag{15}$$

where, B is the batch size, V_i and T_j are normalised image and text embeddings, σ is the sigmoid function, S_{ij} is the similarity matrix value, γ_1 is a reliability

factor, δ_{ij} is the Kronecker delta function (1 for $i = j$, 0 otherwise), and w_{ij} are dynamic weights with enhanced diagonal values based on similarity:

$$w_{ij} = \begin{cases} 1 + \lambda_{diag} \cdot S_{ii}, & \text{if } i = j \\ 1, & \text{otherwise} \end{cases} \tag{16}$$

The soft contrastive loss extends traditional contrastive learning to handle many-to-many relationships:

$$L_{soft} = \frac{1}{2B}(L_{i2t} + L_{t2i}) \tag{17}$$

where, L_{i2t} and L_{t2i} represent the image-to-text and text-to-image directional losses:

$$L_{i2t} = \alpha_1 \cdot L_{i2t}^{pos} + \beta_1 \cdot L_{i2t}^{neg} \tag{18}$$

$$L_{t2i} = \alpha_1 \cdot L_{t2i}^{pos} + \beta_1 \cdot L_{t2i}^{neg} \tag{19}$$

The positive and negative components are defined as:

$$L_{i2t}^{pos} = \sum_{i=1}^{B} \frac{1}{n_i^{pos}} \sum_{j=1}^{B} -\frac{(V_i \cdot T_j^T)}{\text{Temp}_i} \cdot 1[S_{ij} > \theta] \tag{20}$$

$$L_{i2t}^{neg} = \sum_{i=1}^{B} \frac{1}{n_i^{neg}} \sum_{j=1}^{B} \max\left(0, \frac{(V_i \cdot T_j^T)}{\text{Temp}_i} + m_{ij}\right) \cdot 1[S_{ij} < \theta] \tag{21}$$

where, θ is the similarity threshold, n_i^{pos} and n_i^{neg} are the number of positive and negative samples for the i-th image, Temp_i is an adaptive temperature parameter, and m_{ij} is a dynamic margin:

$$m_{ij} = m_{base} + \gamma_1 \cdot (1 - S_{ij}) \cdot 1[S_{ij} < \theta] \tag{22}$$

The adaptive temperature Temp_i is computed based on the contrast between positive and negative similarities:

$$\text{Temp}_i = \text{Temp}_{base} \cdot (1 - 0.3 \cdot \max(0, S_i^{pos} - S_i^{neg})) \tag{23}$$

where, S_i^{pos} and S_i^{neg} are the mean similarities of positive and negative samples for the i-th image, Temp_{base} is the initial temperature parameter.

The overall weighted contrastive loss is formulated as:

$$L_{total}^{WCL} = \alpha_2 \cdot L_{str} + \beta_2 \cdot L_{soft} \tag{24}$$

where, α_2 and β_2 are dynamic weights that adjust based on training progress, with α_2 decreasing from 0.6 to 0.2 and β_2 increasing from 0.4 to 0.8 as training progresses.

Knowledge Distillation Loss Our knowledge distillation architecture employs a comprehensive loss function to facilitate effective knowledge transfer while preserving the semantic relationships essential for medical image-text alignment. The architecture combines cosine similarity loss, Kullback-Leibler (KL) divergence loss, and mean squared error (MSE) loss.

The cosine similarity loss preserves the directional alignment between student and teacher embeddings:

$$L_{mod}^{cos} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \cos\left(\hat{z}_i^{(mod,S)}, \hat{z}_i^{(mod,T)}\right)\right) \tag{25}$$

where, $\hat{z}_i$ represents the normalised feature vectors. S and T seperately represents student vectors and teacher vertors.

The KL divergence loss ensures that the student model learns the distributional characteristics of the teacher embeddings:

$$L_{mod}^{KL} = \tau^2 \cdot D_{KL}\left(p_\tau\left(z^{(mod,S)}\right) \| p_\tau\left(z^{(mod,T)}\right)\right) \tag{26}$$

where, τ is the temperature parameter (default: 2.0), and p_τ represents the softmax-normalised feature distributions.

Additionally, the MSE loss is computed between raw feature vectors to capture absolute differences:

$$L_{mod}^{MSE} = \frac{1}{BD}\sum_{i=1}^{B}\left\|\hat{z}_i^{(mod,S)} - \hat{z}_i^{(mod,T)}\right\|_2^2 \tag{27}$$

The overall knowledge distillation loss is formulated as:

$$L_{total}^{KD} = \alpha_3 \cdot L_{mod}^{MSE} + \beta_3 \cdot L_{mod}^{KL} + \gamma_3 \cdot L_{mod}^{cos} \tag{28}$$

where, α_3, β_3 and γ_3 are weighted parameters (0.33, 0.33, 0.34).

Caption Loss The caption loss in the proposed KCLVA is designed to optimise the generation of radiological reports by training the decoder to predict the next token in the sequence, given the preceding tokens and the aligned image-text features. This loss is computed using teacher forcing and a cross-entropy objective. The caption loss is defined as:

$$L_{caption} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} m_{i,t} \cdot \log p\left(y_{i,t} \mid y_{i,<t}, z_i^{img}, z_i^{txt}\right) \tag{29}$$

where, N is the batch size, T is the sequence length, $y_{i,t}$ is the ground truth token at position t for sample i, $\log p\left(y_{i,t} \mid y_{i,<t}, z_i^{img}, z_i^{txt}\right)$ is the predicted probability of the token $y_{i,t}$ conditioned on all previous tokens $y_{i,<t}$, the image features z_i^{img}, and the text features z_i^{txt}. Additionally, $m_{i,t}$ is the attention mask value, which is 1 if the token is valid and 0 if it is padding.

To implement teacher forcing, the predicted logits are shifted by one position relative to the ground truth labels. Specifically:

$$\text{logits}_{i,t} = \text{decoder}\left(y_{i,<t}, z_i^{img}, z_i^{txt}\right) \tag{30}$$

Here, the predicted logits are shifted to exclude the last token:

$$y_{i,t} = \text{labels}_{i,t+1} \tag{31}$$

Similarly, the ground truth labels are shifted to exclude the first token.

Total Loss Total loss is the combination of caption loss, knowledge distillation loss and caption loss:

$$L_{total} = \alpha_4 \cdot L_{total}^{KD} + \beta_4 \cdot L_{total}^{WCL} + \gamma_4 \cdot L_{caption} \tag{32}$$

where, α_4, β_4 and γ_4 are weighted parameters (0.5, 0.5, 1.0).

4 Experiment

4.1 Dataset

In our experiments, we utilised the widely used benchmark dataset, Indiana University Chest X-ray Collection (IU-Xray), for evaluation. The IU-Xray dataset comprises 7,470 chest X-ray images and 3,955 radiology reports. Each report is divided into three sections: 'Indication', 'Findings', and 'Impression'.

The 'Indication' section details symptoms (e.g., hypoxia) or reasons for the examination (e.g., age); the 'Findings' section lists radiological observations; and the 'Impression' section outlines the final diagnosis. Ideally, a system should generate the 'Findings' and 'Impression' sections, potentially linking them to provide a cohesive diagnostic report. Following previous work [1, 7, 6, 9, 15, 25, 32, 35, 39], we split the dataset into training, validation, and test sets in a 7:1:2 ratio to ensure a fair comparison.

4.2 Experimental Settings

Data Preprocessing Our data preprocessing primarily focuses on text preprocessing, while image preprocessing is managed by the processors of pre-trained image encoders. Specifically, images are resized to 224×224 pixels to standardise input dimensions.

The text preprocessing pipeline for medical reports consists of several sequential steps designed to standardise and normalise the textual data while preserving critical medical information. First, the system employs a comprehensive medical abbreviation dictionary to expand common medical acronyms and abbreviations (e.g., "ct", "copd", "ecg") into their complete forms. Simultaneously, it ensures that the capitalisation of specific medical terms (e.g., "COVID", "COPD") is retained.

Next, the preprocessing workflow applies a series of text cleaning operations. These include the removal of non-medical punctuation using regex patterns, elimination of redundant characters, and normalisation of whitespace and punctuation marks. Privacy-related placeholders (e.g., "XXXX") are systematically removed to ensure compliance with data privacy regulations. Additionally, the text undergoes case normalisation while preserving the integrity of domain-specific terminology. Finally, the process concludes with sentence-level formatting to ensure proper punctuation and maintain semantic coherence.

Implementation Details For the student encoder, we use "vit-base-patch16-224-in21k" [34] as the image encoder and "distilbert-base-uncased" [28] as the text encoder. For the teacher encoder, we select "rad-dino" [23] as the image encoder and "Bio_ClinicalBERT" [2] as the text encoder. The teacher encoder remains frozen during training.

The view-specific attention mechanism is based on multi-head attention, consisting of a single layer with four heads. The output of the student image encoder is processed using one multi-head attention layer, also configured with four heads. For the decoder, we employ two cross-modal transformer layers to fuse the image and text features, followed by four transformer decoder layers for decoding.

The training setup includes a learning rate of 1e-4 with cosine decay, a batch size of 80, a contrastive learning temperature of 0.07, and a knowledge distillation temperature of 2.0. The model is trained for 40 epochs, on the University of Leeds HPC system Aire, using 1 NVIDIA L40S GPU for approximately 5 hours. Additionally, we extract medical terms and construct the similarity matrix, on the University of Leeds HPC system Aire, using 1 NVIDIA L40S GPU and 16×3G CPU cores, which takes approximately 5 hours.

Evaluation Metrics Following the standard evaluation paradigm, we employ the widely-used metrics BLEU [21], METEOR [3], ROUGE-L [14], and CIDEr [31] to assess the quality of generated diagnostic reports.

BLEU evaluates n-gram overlap between the generated and reference texts, capturing precision at different granularities, which is critical for ensuring accurate medical terminology. METEOR complements BLEU by accounting for synonyms, stemming, and word order, providing a more nuanced evaluation of linguistic variations common in medical reports. ROUGE-L emphasises recall by focusing on the longest common subsequence, ensuring that the generated text retains the essential content of the reference. CIDEr, designed for consensus-based evaluation, measures the similarity of the generated text to multiple references using TF-IDF weighting. This makes CIDEr particularly effective for assessing the relevance and informativeness of medical reports.

4.3 Results and Discussion

Result The comparison results are presented in Table 1. Several baseline methods are complex and, in some cases, not open source. Due to limitations in time and computational resources, we were unable to perform multiple runs for these

baselines. Therefore, our comparisons are based on single-run results. Our proposed KCLVA architecture demonstrates superior performance across several evaluation metrics. Specifically, KCLVA achieves the highest BLEU-1 (0.511), BLEU-2 (0.345), BLEU-3 (0.246), METEOR (0.432), and ROUGE-L (0.462) scores, surpassing all other methods in these categories. These results underscore the model's ability to generate captions with improved semantic alignment and greater phrase overlap with the ground truth, producing clinically meaningful and coherent reports. Although KCLVA achieves competitive performance on BLEU-4 (0.173), it is marginally lower than MRCL (0.180), indicating scope for improvement in capturing finer-grained n-gram overlaps. Furthermore, the CIDEr score of KCLVA (0.303) exceeds that of most methods but is lower than METransformer (0.435) and AMLMA (0.381). This suggests that while KCLVA excels in semantic and structural alignment, there is potential to further enhance its ability to capture consensus information across multiple reference reports.

Table 1. Result of comparison on IU-Xray. Comparison results are from test set and the best performance is indicated in bold. * indicates the results are quoted from their published literatures.

Model	BLUE-1	BLUE-2	BLUE-3	BLUE-4	METEOR	ROUGE	CIDEr
R2Gen*[7]	0.470	0.304	0.219	0.165	0.187	0.371	-
R2GenCMN*[?]	0.475	0.309	0.222	0.170	0.191	0.375	-
CMCL*[15]	0.473	0.305	0.217	0.162	0.186	0.378	-
CDGPT2*[1]	0.387	0.245	0.166	0.111	0.164	0.289	0.257
AlignTransformer*[39]	0.484	0.313	0.225	0.173	0.204	0.379	-
Qin and Song*[25]	0.494	0.321	0.235	0.109	0.201	0.384	-
AMLMA*[9]	0.471	0.315	0.231	0.172	0.247	0.376	0.381
MRCL*[35]	0.458	0.324	0.238	**0.180**	0.206	0.369	0.287
METransformer*[32]	0.483	0.322	0.228	0.172	0.192	0.380	**0.435**
Ours(KCLVA)	**0.511**	**0.345**	**0.246**	0.173	**0.432**	**0.462**	0.303

Fig. 2 presents a comparison between the generated reports and the reference reports. Two examples, including one AP view and one LTA view, are showcased. Similar words between generated reports and reference reports are highlighted in green text, demonstrating the system's ability to generate clinically relevant observations that align closely with the reference reports

Discussion Although our KCLVA model demonstrates promising performance on standard metrics, several limitations remain. The reliability and scalability of the knowledge base are critical—if the knowledge base is insufficient, accurate extraction of medical terms becomes challenging and may introduce noise into the training process. The model's performance is closely tied to the quality of UMLS term extraction, a process that is not only time-consuming—requiring up to 5 hours for preprocessing on the IU-Xray dataset—but also prone to instability and inaccuracy, particularly in the absence of sufficient clinical guidance. Additionally, the model may focus excessively on view-specific diagnoses while

neglecting the global context. Furthermore, the model is constrained by long-tail data distributions in the dataset, resulting in reduced accuracy when generating reports for rare or uncommon pathologies. In such cases, it often defaults to descriptions associated with more prevalent conditions, reflecting a tendency toward majority class bias when confronted with unusual presentations.

To address these challenges, future work could focus on enhancing the reliability and scalability of the knowledge base, as well as improving medical term extraction through closer collaboration with medical experts and optimizing both UMLS search and similarity computation. Specifically, to mitigate the impact of long-tail data distributions—such as the overrepresentation of normal conditions like "no pleural effusions" and "lungs are clear" in the dataset compared to rarer conditions—incorporating explicit medical knowledge graphs may enhance the model's reasoning about rare conditions. Such approaches could help balance the training environment and improve the model's ability to generate accurate reports for both common and uncommon pathologies.

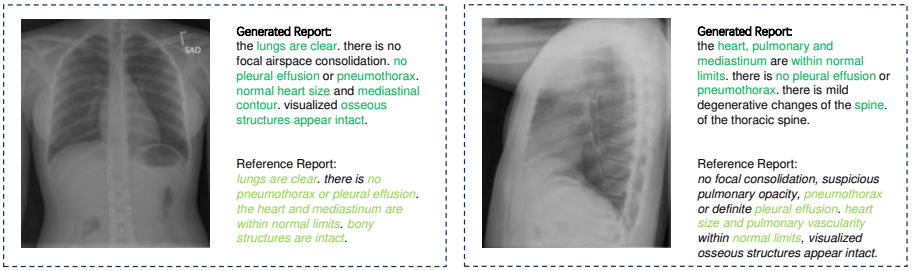


Fig. 2. Two examples of the generated reports and their comparison with reference reports for chest X-ray images. The figure shows one anteroposterior case and one lateral case, each containing a chest X-ray image with its corresponding AI-generated report (top) and the radiologist's reference report (bottom). The color coding (dark green for generated reports and light green for reference reports) highlights the similar words between the AI-generated report and the radiologists' report.

4.4 Ablation Study

We conducted an ablation study to evaluate the effectiveness of VA and the WCL by replacing the WCL with one-to-one Noise Contrastive Estimation (InfoNCE) [20] loss and adding or removing the VA. Notably, the VA is always utilised in conjunction with the UMLS-based term extractor. The extractor is not employed when VA is omitted. In this study, each architecture is trained five times [18] to validate the effect of the proposed WCL and VA, using the same split of training, validation, and test datasets. The results are presented in Table 2.

The results demonstrate that the combination of WCL and VA achieves the best performance across all metrics, with significant improvements in BLEU-4, METEOR, ROUGE, and CIDEr scores, underscoring their complementary benefits. When comparing WCL with InfoNCE, WCL+VA consistently outperforms InfoNCE+VA across all evaluation criteria, reaffirming the advantages of WCL in multimodal learning tasks. However, without VA integration, the performance

gap between WCL and InfoNCE narrows, with WCL-VA showing only modest improvements over InfoNCE-VA in most metrics. This suggests that the benefits of WCL are amplified when paired with attention mechanisms.

Interestingly, InfoNCE-VA slightly outperforms InfoNCE+VA on several metrics (BLEU-1, BLEU-2, and METEOR), indicating potential challenges in integrating InfoNCE with VA effectively. This highlights the importance of designing loss functions tailored to structured attention mechanisms. In contrast, WCL demonstrates consistent improvements when combined with VA, validating that many-to-many relationship learning provides significant advantages over traditional one-to-one contrastive approaches in multimodal contexts. The results further emphasize the robustness of WCL+VA in achieving best performance.

Table 2. Ablation Study Results. "-VA" indicates that VA is not used, "+VA" indicates that VA is utilised, "InfoNCE" refers to replacing the WCL with InfoNCE, and "WCL" denotes the use of the proposed WCL. The best performance is indicated in bold. Each evaluation metrics is composed of "five-time result average ± five-time result standard deviation", representing the range of evaluation metrics.

Model	BLUE-1	BLUE-2	BLUE-3	BLUE-4	METEOR	ROUGE	CIDEr
InfoNCE-VA	0.455 ±0.024	0.285 ±0.017	0.195 ±0.013	0.137 ±0.011	0.387 ±0.019	0.405 ±0.003	0.258 ±0.021
InfoNCE+VA	0.460 ±0.008	0.286 ±0.012	0.196 ±0.012	0.138 ±0.010	0.392 ±0.012	0.406 ±0.012	0.244 ±0.024
WCL-VA	0.462 ±0.009	0.288 ±0.004	0.196 ±0.002	0.139 ±0.002	0.398 ±0.015	0.401 ±0.011	0.268 ±0.050
WCL+VA	**0.487 ±0.022**	**0.315 ±0.022**	**0.219 ±0.018**	**0.153 ±0.013**	**0.418 ±0.012**	**0.429 ±0.020**	**0.278 ±0.042**

5 Conclusions

In this paper, we propose a novel encoder-decoder-based architecture, KCLVA, for chest X-ray (CXR) report generation. By leveraging UMLS, view-specific attention, and a weighted contrastive loss, our model effectively aligns multi-view CXR images with their corresponding diagnostic reports. The proposed architecture emulates the diagnostic process of radiologists by focusing on view-specific features while preserving the global context. Experimental results on the IU-Xray dataset demonstrate the effectiveness of the KCLVA architecture. Nonetheless, KCLVA is constrained by the long-tail distribution and its reliance on the medical knowledge base, as it is highly dependent on the time-consuming process of term extraction and the construction of the similarity matrix. To address these issues, we plan to construct a more scalable knowledge resource and collaborate with radiologists to supervise preprocessing. Additionally, we aim to optimize preprocessing to enhance scalability. For future work, we plan to further assess the generalizability of our model by evaluating it on larger and more diverse datasets, such as MIMIC-CXR.

References

1. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., Fahmy, A.: Automated radiology report generation using conditioned transformers. Informatics in Medicine Unlocked **24**, 100557 (2021)
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
5. Cao, W., Zhang, J., Xia, Y., Mok, T.C., Li, Z., Ye, X., Lu, L., Zheng, J., Tang, Y., Zhang, L.: Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11238–11247 (2024)
6. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
7. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
9. Gajbhiye, G.O., Nandedkar, A.V., Faye, I.: Translating medical image to radiological report: Adaptive multilevel multi-attention approach. Computer Methods and Programs in Biomedicine **221**, 106853 (2022)
10. Hosseinzadeh, H.: Deep multi-view feature learning for detecting covid-19 based on chest x-ray images. Biomedical Signal Processing and Control **75**, 103595 (2022)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
12. Kim, D.: Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2702–2710 (2023)
13. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
15. Liu, F., Ge, S., Zou, Y., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. arXiv preprint arXiv:2206.14579 (2022)
16. Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. arXiv preprint arXiv:2106.06965 (2021)

17. Liu, Z., Zhu, Z., Zheng, S., Zhao, Y., He, K., Zhao, Y.: From observation to concept: A flexible multi-view paradigm for medical report generation. IEEE Transactions on Multimedia **26**, 5987–5995 (2023)
18. Lu, P., Wang, C., Hagenah, J., Ghiasi, S., Zhu, T., Thwaites, L., Clifton, D.A., et al.: Improving classification of tetanus severity for patients in low-middle income countries wearing ecg sensors by using a cnn-transformer network. IEEE Transactions on Biomedical Engineering **70**(4), 1340–1350 (2022)
19. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669 (2019)
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
22. Paul, A., Shen, T.C., Lee, S., Balachandar, N., Peng, Y., Lu, Z., Summers, R.M.: Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training. IEEE Transactions on Medical Imaging **40**(10), 2642–2655 (2021)
23. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Rad-dino: Exploring scalable medical image encoders beyond text supervision. arXiv preprint arXiv:2401.10815 (2024)
24. Prabhakar, C., Sekuboyina, A., Paetzold, J.C., Li, H., Amiranashvili, T., Kleesiek, J., et al.: Improving generalized zero-shot learning for multi-labelchest x-ray classification using knowledge graphs (2021)
25. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 448–458 (2022)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
27. Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M.: Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint arXiv:1804.07839 (2018)
28. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
29. Speets, A.M., van der Graaf, Y., Hoes, A.W., Kalmijn, S., Sachs, A.P., Rutten, M.J., Gratama, J.W.C., van Swijndregt, A.D.M., Mali, W.P.: Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. British Journal of General Practice **56**(529), 574–578 (2006)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
32. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11558–11567 (2023)

33. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
34. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
35. Wu, X., Li, J., Wang, J., Qian, Q.: Multimodal contrastive learning for radiology report generation. Journal of Ambient Intelligence and Humanized Computing **14**(8), 11185–11194 (2023)
36. Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., Hsu, C.N.: Weakly supervised contrastive learning for chest x-ray report generation. arXiv preprint arXiv:2109.12242 (2021)
37. Yang, S., Niu, J., Wu, J., Liu, X.: Automatic medical image report generation with multi-view and multi-modal attention mechanism. In: International Conference on Algorithms and Architectures for Parallel Processing. pp. 687–699. Springer (2020)
38. Yang, Y., Yu, J., Jiang, H., Han, W., Zhang, J., Jiang, W.: A contrastive triplet network for automatic chest x-ray reporting. Neurocomputing **502**, 71–83 (2022)
39. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 72–82. Springer (2021)
40. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
41. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
42. Zhu, X., Feng, Q.: Mvc-net: Multi-view chest radiograph classification network with deep fusion. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 554–558. IEEE (2021)