

KiUT: Knowledge-injected U-Transformer for Radiology Report Generation

Zhongzhen Huang^{1,2}, Xiaofan Zhang^{1,2} ✉, Shaoting Zhang^{2,3}

¹Shanghai Jiao Tong University ²Shanghai AI Laboratory ³SenseTime Research

{huangzhongzhen, xiaofan.zhang}@sjtu.edu.cn, zhangshaoting@pjlab.org.cn

Abstract

Radiology report generation aims to automatically generate a clinically accurate and coherent paragraph from the X-ray image, which could relieve radiologists from the heavy burden of report writing. Although various image caption methods have shown remarkable performance in the natural image field, generating accurate reports for medical images requires knowledge of multiple modalities, including vision, language, and medical terminology. We propose a Knowledge-injected U-Transformer (KiUT) to learn multi-level visual representation and adaptively distill the information with contextual and clinical knowledge for word prediction. In detail, a U-connection schema between the encoder and decoder is designed to model interactions between different modalities. And a symptom graph and an injected knowledge distiller are developed to assist the report generation. Experimentally, we outperform state-of-the-art methods on two widely used benchmark datasets: IU-Xray and MIMIC-CXR. Further experimental results prove the advantages of our architecture and the complementary benefits of the injected knowledge.

1. Introduction

Radiology images (e.g., chest X-ray) play an indispensable role in routine diagnosis and treatment, and the radiology reports of images are essential in facilitating later treatments. Getting a hand-crafted report is a time-consuming and error-prone process. Given a radiology image, only experienced radiologists can accurately interpret the image and write down corresponding findings. Therefore, automatically generating high-quality radiology reports is urgently needed to help radiologists eliminate the overwhelming volume of radiology images. In recent years, radiology report generation has attracted much attention in the deep learning and medical domain. The encoder-decoder architecture inspired by neural machine translation [34] has been widely adopted by most existing methods [17, 24, 40, 42].

✉ Corresponding Author.

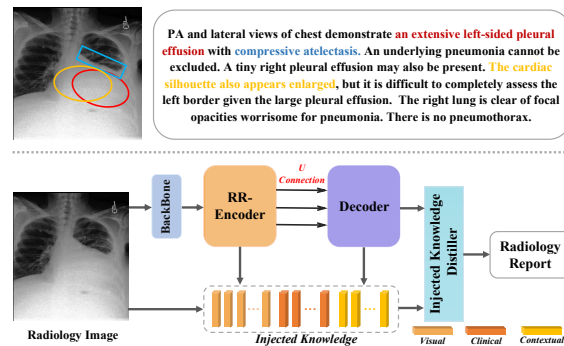


Figure 1. A transformer architecture with U-connection is adopted to generate reports from radiology images. The process involves injecting and distilling visual, clinical, and contextual knowledge. The color labels in the image and report represent the different abnormal regions and their corresponding description, respectively.

With the recent advent of the attention mechanism, the architecture’s capability is greatly ameliorated.

Despite the remarkable performance, these models restrained themselves in the methodology of image caption [6, 7, 36, 39, 43], and suffer from such data biases: 1) the normal cases dominate the dataset over the abnormal cases; 2) the descriptions of normal regions dominate the entire report. Recently, some methods have been proposed to either alleviate case-level bias by utilizing posterior and prior knowledge [27] or relieve the region-level bias by distilling the contrastive information of abnormal regions [28].

Thus, in the medical field’s cross-modal task, a model needs to not only capture visual details of every abnormal region but also consider the interaction between the visual and textual modalities among different levels. Moreover, external clinical knowledge is required to achieve the radiologist-like ability in radiology image understanding and report writing. The external knowledge, e.g., the clinical entities and relationships, could be pre-defined by experts or mined from medical documents. However, directly adopting the knowledge brings inconsistencies due to the heterogeneous context embedding space [21]. And too complex knowledge may be prone to distract the visual encoder and divert the representation [27].

Instead of using external knowledge to augment the feature extraction like previous approaches [27, 44], we propose to introduce the injected knowledge in the final decoding stage. A graph with the clinical entities, *i.e.*, symptoms and their relationships, is constructed under the guidance of professional doctors. These entities have homogeneous embedding space with the training corpus, and this signal could be injected smoothly with visual and contextual information. We further design the Injected Knowledge Distiller on top of the decoder to distill contributive knowledge from visual, contextual, and clinical knowledge.

Following these premises, we explore a novel framework dubbed as *Knowledge-injected and U-Transformer* (KiUT) to achieve the radiologist-like ability to understand the radiology images and write reports. As Fig. 1 shows, it consists of a Region Relationship Encoder and Decoder with U-connection architecture and Injected Knowledge Distiller.

Our contributions can be summarized as follows:

- We propose a novel model following the encoder-decoder architecture with U-connection that fully exploits different levels of visual information instead of only one single input from the visual modality. In our experiments, the U-connection schema presents improvement not only in radiology report generation but also in the natural image captioning task.
- Our proposed model injects clinical knowledge by constructing a symptom graph, combining it with the visual and contextual information, and distilling them when generating the final words in the decoding stage.
- The Region Relationship Encoder is developed to restore the extrinsic and intrinsic relationships among image regions for extracting abnormal region features, which are crucial in the medical domain.
- We evaluate our approach on two public radiology report generation datasets, IU-Xray [8] and MIMIC-CXR [18]. *KiUT* achieves state-of-the-art performance on the two benchmark datasets.

2. Related Work

2.1. Image Caption

Generating radiology reports is similar to the image caption task, which aims to describe the content of a given natural image. Based on the encoder-decoder architecture, the development of image caption models [16,30,33,37,39] has advanced in leaps and bounds.

With the good performance of Transformer [35] in both computer vision and natural language processing field, a broad collection of transformer based methods have been explored to ameliorate the performance of Image Caption. A combined bottom-up and top-down attention mechanism was introduced by [1] to calculate the relationships between objects and other salient image regions when extracting vi-

sual representation. [7] proposed a meshed-memory transformer that represents visual features incorporating the relationships between image regions. In the work [43], Zhang *et al.* enhanced visual representations in the attention module with relative geometry features and used adaptive attention to predict visual and non-visual words. *GRIT* [31] adopts a transformer architecture to fuse the objects' feature and contextual features for better captioning. Although these methods are effective, they still need to take full advantage of the visual information and consider the interaction between the encoder and decoder. To explore the relationship between these two modalities, [7] tried complex mesh-like connectivity at the decoding stage. In this paper, we try a simple and reasonable schema to establish the connection between the encoder and decoder.

2.2. Radiology Report Generation

Radiology reports contain multiple descriptive sentences about radiology images. As the transfer and extension of Image Caption in the medical field, radiology report generation not only puts forward higher requirements on the length of generated reports but also presents greater challenges on the accuracy of long contextual descriptions. Much work has made some advancements in this task. [23] employed a hierarchical decision making procedure and generated normal and abnormal sentences by multi-agent systems, respectively. Liu *et al.* [29] presented a hierarchical generation framework which first predicts the related topics of the input, then generates sentences according to the topics. For the latter studies, transformer [35] has been successfully applied as the model design rationale. [13] developed a transformer-based architecture and used attention to localize important regions in the image for generation. Chen *et al.* [3] improved the performance by incorporating the transformer based model with memory matrix which is designed to align cross modal features.

Recently, some work [20–22, 27, 38, 44] began to explore assisting report generation with additional knowledge. Wang *et al.* [38] utilized a shared cross-modal prototype matrix as external knowledge to record the cross-modal prototypes and embed the cross-modal information for the reports generation. [44] designed a pre-constructed medical graph based on prior knowledge from chest findings, which allows dedicated feature learning for each disease finding. The graph improves the models' capability to understand medical domain knowledge. Inspired by the idea, Liu *et al.* explored and distilled posterior and prior knowledge in the work [27], where the posterior knowledge is visual information and the prior knowledge is the medical graph and retrieval reports. Instead of applying knowledge for visual feature extraction, we distill knowledge in the decoding stage. The report generated by our model is based on the distilled knowledge from multiple modalities.

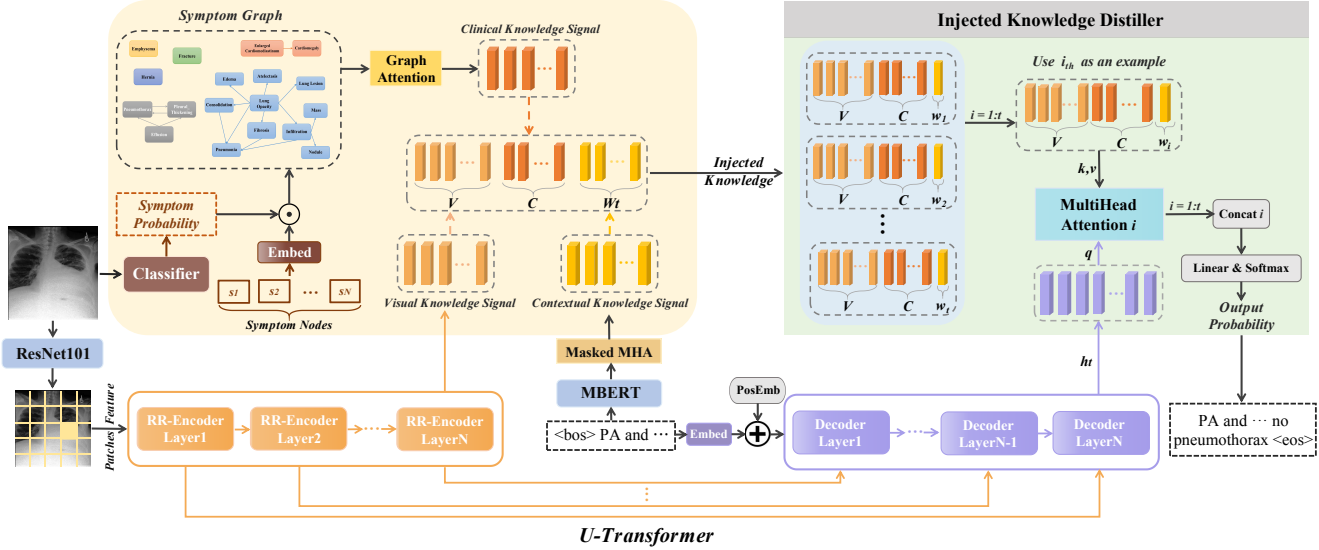


Figure 2. The overall architecture of *KiUT*: 1) U-Transformer with U-connection between Visual RR-Encoder and Decoder, 2) Injected Knowledge Distiller that handles three types of injected knowledge. There are seriatim elaborate descriptions of these modules in Section 3.

3. Methodology

In the radiology report generation task, given a 2D radiology image I , the model is required to interpret the image and generate a descriptive radiology report $R = \{y_1, y_2, \dots, y_{N_R}\}$, where y_i is the word token of the report and N_R is the length of the report. The entire recursive generation process can be formulated as follows:

$$p(R | I) = \prod_{t=1} p(y_{t+1} | y_1, \dots, y_t, I) \quad (1)$$

And the model is generally optimized by minimizing the cross-entropy loss:

$$L_{CE}(\theta) = - \sum_{i=1}^N \log(p_{\theta}(y_n^* | y_{1:n-1}^*)) \quad (2)$$

where $R^* = \{y_1^*, y_2^*, \dots, y_{N_R}^*\}$ is the ground truth report.

In this section, we will introduce the framework of the proposed *KiUT*. Fig. 2 shows the overall architecture of our proposed *KiUT*. Our model can be conceptually divided into three core components: Cross-modal U-Transformer, Injected Knowledge Distiller, and Region Relationship Encoder. We will introduce their details in turn.

3.1. Cross-modal U-Transformer

For cross-modal tasks, it is particularly important to obtain proper interaction between different modalities. As shown in Fig. 3, compared with the previous work using only the last output of the encoder, or proposing a meshed connection [7], we design a U-transformer architecture with much fewer parameters than the meshed con-

nection. Specifically, we construct a U-connection that connects the layers of the encoder and decoder.

The visual encoder captures details in X-ray images by the multi-head self-attention that naturally aggregates task-relevant features. The decoder aggregates the visual features throughout all layers via U-connection. And the output features of the last decoder layer are sent to the Injected Knowledge Distiller for generating the words in the report.

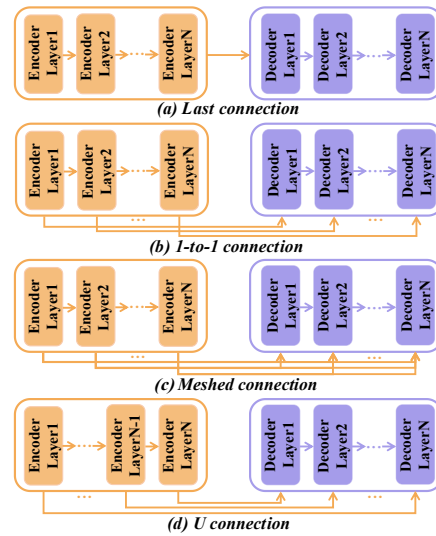


Figure 3. Connection schemas between the encoder and decoder.

Assume that the number of layers of the proposed encoder and decoder is N . The output of each layer of the encoder can be formulated as $\{\tilde{x}_1, \dots, \tilde{x}_i, \dots, \tilde{x}_N\}$, where \tilde{x}_i is defined as the i -th layer's output of our Visual RR-

Encoder (Sec. 3.3). In our proposed architecture, the output of the i -th encoder layer will be input to the $(N - i + 1)$ -th decoder layer. The specific process can be formulated as:

$$\{\tilde{x}_1, \dots, \tilde{x}_N\} = \text{RR-Encoder}(I) \quad (3)$$

$$\hat{x}_i = \tilde{x}_{N-i+1} \quad (4)$$

where \hat{x}_i is the input of the i -th decoder layer.

After receiving the visual features $\{\hat{x}_1, \dots, \hat{x}_N\}$ from the encoder, the decoder generates the hidden state h_t to predict the current word y_{t+1} through the word sequence features \hat{w}_t in the decoding step $t+1$. Given the generated words $y_{1:t} = \{y_1, \dots, y_t\}$, the word sequence features are obtained by words embeddings w_t combining with its position encodings e_t . The process of the decoder is as follows:

$$e_t = \text{PosEncoding}(y_{1:t}) \quad (5)$$

$$\hat{w}_t = w_t + e_t \quad (6)$$

$$h_t^i = \begin{cases} \text{DecoderLayer}_i(\hat{w}_t, \hat{x}_1) & i = 1 \\ \text{DecoderLayer}_i(h_t^{i-1}, \hat{x}_i) & i > 1 \end{cases} \quad (7)$$

where the DecoderLayer_i and h_t^i represent the i -th DecoderLayer and its output, the final hidden state $h_t = h_t^N$. DecoderLayer and PosEncoding follows the operations of *Decoder* and *Position encoding* in [35].

3.2. Injected Knowledge Distiller

Radiologists draw on different aspects of medical knowledge when writing reports. To make the process of report generation conformance to the real scenario and generate more accurate reports, our model incorporates injected knowledge signals from three aspects, *i.e.*, visual knowledge, contextual knowledge, and clinical knowledge.

Visual knowledge signal contains the information of the radiology image I . \tilde{X} is the output of the encoder's last layer, namely $\tilde{X} = \tilde{x}_N$.

Contextual knowledge signal is the information of the generated words extracted from *MBert* with a masked attention module. Specifically, to endow *MBert* with partial medical language knowledge, we adopt a pre-trained *Bert* model [9] and finetune it on the reports in train split of the specific dataset [8, 18]. And a MaskedMHA [35] module is applied after *MBert* to obtain linguistic features of the generated report words. The process of extracting contextual knowledge signal can be defined as:

$$\tilde{W}_t = \text{MaskedMHA}(\text{MBert}(y_{1:t}) + e_t) \quad (8)$$

where $\tilde{W}_t \in \mathbb{R}^{t \times d}$ and $y_{1:t}$ is the generated words sequence.

Clinical knowledge signal is essential to the radiology report generation task. Imaging exam is scheduled for a specific purpose in the medical field. For example, the chest X-ray is a part of the physical exam that could show the

size, shape, and location of the heart, lungs, bronchi, pulmonary arteries, *etc.* Therefore, we could design a clinical symptom graph to inject real-world medical knowledge into the model. The graph shown in Fig. 4 is developed according to the professional perspective of radiology images, taking into account symptoms correlation, symptom characteristics, occurrence location, *etc.* The feature of each node $sf_i \in \mathbb{R}^d$ in the graph is the word embedding of each symptom derived from the *MBert*.

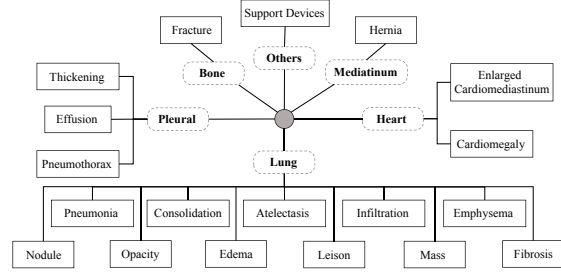


Figure 4. The symptoms graph is based on the correlation, characteristics and occurrence location of symptoms.

For every radiology image I , the probability distribution sp_i of each symptom in the graph is calculated with the pre-trained *classification model* [5]. Then, the symptom graph G is initialized according to the distribution, and the graph attention mechanism (GAT) is utilized to calculate the clinical knowledge signal \tilde{C} . The process is as follows:

$$sg_i = sp_i \odot sf_i, g_i \in G \quad (9)$$

$$\tilde{C} = \text{GAT}(G) \quad (10)$$

where g_i represents i th node of the symptom graph G .

The clinical knowledge signal indicates the possible findings that may need to be written in the report. And these findings reflect the radiologists' first impression of the image based on their professional knowledge.

Knowledge Distiller To make our model achieve the radiologist-like ability to generate reports and align with the real scenario, we propose an Injected Knowledge Distiller to distill useful information from the above knowledge signals. The process is formally defined as follows:

$$\tilde{F}_i = [\tilde{X}, \tilde{C}, \tilde{w}_i] \quad i \in [1, t], \tilde{w}_i \in \tilde{W}_t \quad (11)$$

$$\text{MHA}_i(h_t, \tilde{F}_i) = \text{Softmax}\left(\frac{QK_i^T}{\sqrt{d_n}}\right) V_i \quad (12)$$

$$Q = h_t W^Q, K_i = \tilde{F}_i W^K, V_i = \tilde{F}_i W^V \quad (13)$$

$$\tilde{h}_t = \text{Concat}(\text{MHA}_1(h_t, \tilde{F}_1), \dots, \text{MHA}_t(h_t, \tilde{F}_t)) \quad (14)$$

$$y_{t+1} \sim p_{t+1} = \text{Softmax}\left(\tilde{h}_t W_p + b_p\right) \quad (15)$$

where $\tilde{w}_i \in \tilde{W}_t$, W^Q , W^K , W^V , W_p and b_p are learnable parameters. $y_{t+1} \sim p_{t+1}$ is the probability of the generated word at step $t + 1$.

3.3. Region Relationship Encoder

To get the visual representation, the first step is to extract features from the radiology image. Following [3, 4, 38], we adopt pre-trained convolutional neural networks to extract the basic visual feature X from a radiology image I . Generally, the image is decomposed into S patches of equal size by flattening the extracted feature by row. In brief, $X = \{x_i\}_{i=1}^S$ is the source input for the subsequent modules, and $x_i \in \mathbb{R}^d$.

After getting a sequence of the visual features X extracted from radiology image regions, it is crucial to explore the extrinsic and intrinsic relationships among these regions for better visual representations, which are important for understanding radiology images. For example, the extrinsic relationship: *the heart is located below the left lung* and the intrinsic relationship: *enlarged cardiomeastinum is usually associated with some left lung imaging manifestations*. Although *self-attention* [35] can be used to model the relationship among image regions, it can only calculate the similarity between image region features [7]. Therefore, we propose a region relationship encoder to explore the extrinsic and intrinsic relationships among the regions.

For the flattened image region features, the loss of the extrinsic relationships is inevitable, *e.g.*, the spatial information. Therefore, we incorporate relative geometry information into *self-attention* to take into account the extrinsic relationships among regions. We first calculate the center coordinate (x_i, y_i) , width w_i , and height h_i of a region i , the specific calculation as follows:

$$(x_i, y_i) = \left(\frac{x_i^{\min} + x_i^{\max}}{2}, \frac{y_i^{\min} + y_i^{\max}}{2} \right) \quad (16)$$

$$w_i = (x_i^{\max} - x_i^{\min}) + 1 \quad (17)$$

$$h_i = (y_i^{\max} - y_i^{\min}) + 1 \quad (18)$$

where (x_i^{\max}, y_i^{\max}) and (x_i^{\min}, y_i^{\min}) are the position coordinate of the upper left corner and the lower right corner of the grid i , respectively. Following the computation of region relative geometry features in [10, 12], the relative geometry relationship of regions can be calculated by:

$$rg_{ij} = \left(\log \left(\frac{|x_i - x_j|}{w_i} \right), \log \left(\frac{|y_i - y_j|}{h_i} \right), \log \left(\frac{w_i}{w_j} \right), \log \left(\frac{h_i}{h_j} \right) \right)^T \quad (19)$$

where $rg \in \mathbb{R}^{S \times S \times 4}$. And the extrinsic relationship ER_{ij} between region i and j can be obtained by:

$$ER_{ij} = \text{ReLU}(w_g^T FC(rg_{ij})) \quad (20)$$

where w_g is a learnable parameter, FC is a fully-connected layer and ReLU is an activation for zero trimming. $ER \in \mathbb{R}^{S \times S}$ is the representation of the extrinsic relationship.

To represent the intrinsic relationship among regions, we use additional learnable matrices and extend the set of keys and values in *self-attention* with them [7]. This operation is defined as:

$$Keys, Values = W_k X, W_v X \quad (21)$$

$$K_I = [Keys, M_k], V_I = [Values, M_v] \quad (22)$$

where M_k and M_v are learnable matrices with M_N rows, X is the input to *self-attention*, W_k and W_v are learnable parameters. $[\cdot, \cdot]$ stands for concatenation operation.

Based on the above, to compensate for the region relationship in the visual encoder, we propose a region relationship augmented self-attention (\mathcal{RRSA}) in our encoder. The \mathcal{RRSA} is formally defined as follows:

$$\mathcal{RRSA}(Q, K_I, V_I) = \text{Softmax} \left(\frac{QK_I^T}{\sqrt{d_n}} + ER_I \right) V_I \quad (23)$$

$$Q = W_q X, ER_I = [ER, M_I] \quad (24)$$

where W_v and M_I are learnable parameters. The visual features $\{\tilde{x}_1, \dots, \tilde{x}_N\}$ are obtained by the \mathcal{RRSA} based region relationship encoder.

4. Experiment

4.1. Dataset

We conduct experiments to evaluate the effectiveness of the proposed $KiUT$ on two widely used medical report generation benchmarks, *i.e.*, IU-Xray and MIMIC-CXR.

IU-Xray [8] from Indiana University is a relatively small but publicly available dataset containing 7,470 chest X-ray images and 3,955 radiology reports. For the majority of reports, there are frontal and lateral radiology images. Following the previous work [3, 4, 27, 38], we first exclude the samples without "Findings" section and follow the widely-used splits to divide the dataset into train (70%), validation (10%) and test (20%) sets.

MIMIC-CXR [18] provided by the Beth Israel Deaconess Medical Center is a recently released large-scale dataset. The dataset includes 377,110 chest X-ray images and 227,835 reports. For a fair comparison, we adopt the official data splits (*i.e.*, 70%/10%/20% for train/validation/test set). Thus, there are 368,960 in the training set, 2,991 in the validation set, and 5,159 in the test set.

For the reports of these two datasets, we preprocess them by tokenizing and converting all tokens to lower cases and removing the special tokens like digital, non-alphanumeric characters, etc. Finally, we remove the tokens whose frequency of occurrence is less than a specific threshold from the remaining tokens to construct our vocabulary. Fig. 5 shows the statistical visualization of all datasets in terms of the size of vocabulary and the average length of reports.

Table 1. Performance comparisons of the proposed KiUT with existing methods on the test sets of MIMIC-CXR and IU-Xray datasets with respect to NLG and CE metrics. The best values are highlighted in bold.

Dataset	Method	NLG Metric						CE Metric		
		Bleu1	Bleu2	Bleu3	Bleu4	Meteor	Rouge_L	Precision	Recall	F1
IU-Xray	SentSAT+KG [44]	0.441	0.291	0.203	0.147	—	0.367	—	—	—
	R2GenCMN [3]	0.470	0.304	0.219	0.165	0.187	0.371	—	—	—
	PPKED [27]	0.483	0.315	0.224	0.168	0.190	0.376	—	—	—
	AlignTrans [41]	0.484	0.313	0.225	0.173	0.204	0.379	—	—	—
	Contrastive [28]	0.492	0.314	0.222	0.169	0.193	0.381	—	—	—
	XPRONET [38]	0.525	0.357	0.262	0.199	0.220	0.411	—	—	—
	Ours	0.525	0.360	0.251	0.185	0.242	0.409	—	—	—
MIMIC CXR	Up-Down [1]	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
	Att2in [33]	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
	R2GenCMN [3]	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	PPKED [27]	0.360	0.224	0.149	0.106	0.149	0.284	—	—	—
	Contrastive [28]	0.350	0.219	0.152	0.109	0.151	0.283	0.352	0.298	0.303
	AlignTrans [41]	0.378	0.235	0.156	0.112	0.158	0.283	—	—	—
	XPRONET [38]	0.344	0.215	0.146	0.105	0.138	0.279	—	—	—
	Ours	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321

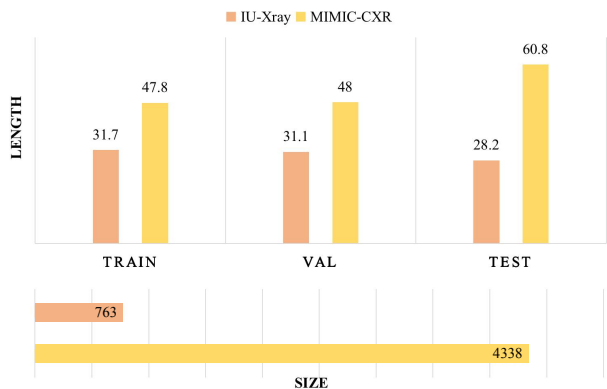


Figure 5. The size of vocabulary and the average length of reports for two datasets.

4.2. Implementation Details

Following the previous work [3, 4, 38], we adopt the ResNet-101 [11] pretrained on the ImageNet [19] to extract image region features. The obtained features are further projected into 512-dimension in the shape of 7×7 , *i.e.* S is 49 and d is 512. To ensure consistency with the existing methods setting [3, 27, 38], we utilize paired images of a patient as the input for IU-Xray and one image for MIMIC-CXR. For the encoder-decoder architecture, we follow the implementation of a Transformer-based model and keep the inner structure and parameters untouched. To facilitate the method, we use *XRayVison* models, a pre-trained DenseNet-121 [14] on radiology images, to generate probabilities for the mentioned symptoms graph.

4.3. Metrics

To gauge the performance, we employ the widely-used natural language generation (NLG) metrics and clinical efficacy (CE) metrics. We adopt the standard evaluation protocol to calculate the captioning metrics: BLEU [32], METEOR [2] and ROUGE-L [25]. And for the clinical efficacy, we apply CheXpert [15] to label the generated reports and compare the labeling results with ground truths in 14 categories of diseases through precision, recall and F1.

4.4. Comparison with state-of-the-art

To demonstrate the effectiveness, we compare the performances of our model with a wide range of state-of-the-art models on the MIMIC-CXR and IU-Xray. Tab. 1 shows the comparison results on both NLG and CE metrics. The models we compare to include Up-Down [1], R2GenCMN [3], PPKED [27], XPRONET [38], *et al.* As shown in Tab. 1, our KiUT outperforms state-of-the-art methods across all metrics on the MIMIC-CXR. Compared to the methods not specially designed for the medical domain [1, 3, 33], our model shows a notable improvement. Furthermore, the result validates that our knowledge injection strategy, *i.e.*, introducing a symptom graph in the final decoding step, performs better than those methods that are specially designed for radiology report generation [27, 28, 38, 41]. Moreover, the superior clinical efficacy scores that measure the accuracy of generated reports for clinical abnormalities demonstrate that our approach can produce higher-quality descriptions for clinical abnormalities.

Our model also significantly outperforms all the other methods in terms of the evaluation metrics on the IU-

Table 2. Ablation study of our method on the MIMIC-CXR dataset, which includes the RR-Encoder and IK-Distiller.

Dataset	RR-Encoder		IK-Distiller		Metric					
	ER	IR	Contextual	Clinical	Bleu1	Bleu2	Bleu3	Bleu4	Meteor	Rouge_L
MIMIC-CXR			✓	✓	0.373	0.224	0.148	0.106	0.137	0.278
		✓	✓	✓	0.382	0.231	0.153	0.110	0.143	0.278
	✓		✓	✓	0.379	0.230	0.154	0.110	0.144	0.281
	✓	✓			0.337	0.207	0.139	0.099	0.132	0.273
	✓	✓	✓		0.361	0.218	0.146	0.101	0.136	0.275
	✓	✓		✓	0.371	0.224	0.149	0.108	0.141	0.278
	✓	✓	✓	✓	0.393	0.243	0.159	0.113	0.160	0.285

Xray datasets, except some benchmarks on the IU-Xray are slightly inferior to XPRONET [38]. This could be partly explained by XPRONET adopting cross-modal prototypes to record the information following [3]. The cross-modal module is easier to learn informative features in the small dataset IU-Xray, but fails to learn adequate information in the MIMIC-CXR whose size is almost 100 times larger than IU-Xray. As Fig. 5 shows, the length of reports in MIMIC-CXR is longer than that in IU-Xray. This is further reflected in our proposed model has better generalizability.

Table 3. Quantitative analysis for the U-connection structure. The experiments conducted on \mathcal{M}^2 Transformer, RSTNet, R2GenCMN (R2CMN) and our KiUT. RSTNet and R2CMN adopt the last connection schema.

Dataset	Method	Bleu1	Bleu4	Meteor	Rouge_L
COCO test [26]	\mathcal{M}^2 Trans [7]	0.808	0.391	0.292	0.586
	\mathcal{M}^2 Trans ^{1-to-1}	0.803	0.382	0.289	0.582
	\mathcal{M}^2 Trans ^U	0.808	0.391	0.289	0.583
	RSTNet [43]	0.811	0.393	0.294	0.588
	RSTNet ^{1-to-1}	0.809	0.398	0.288	0.581
	RSTNet ^U	0.814	0.403	0.293	0.591
MIMIC CXR	R2CMN [3]	0.353	0.103	0.142	0.277
	R2CMN ^{1-to-1}	0.356	0.101	0.141	0.273
	R2CMN ^U	0.363	0.106	0.142	0.276
	KiUT ^{last}	0.372	0.109	0.142	0.279
	KiUT ^{1-to-1}	0.380	0.110	0.143	0.281
	KiUT ^U	0.393	0.113	0.160	0.285

4.5. Ablation Studies

To fully investigate the contribution of our proposed Region Relationship Encoder(RR-Encoder), Injected Knowledge Distiller(IK-Distiller), and the U-connection schema. We conduct a quantitative analysis to compare each component. The main results are shown in Tab. 2 and Tab. 3.

Effect of RR-Encoder We start from a basic encoder without the extrinsic relationship (ER) and the intrinsic relationship (IR). Then we add ER and IR, respectively. Comparing the results in Tab. 2, both ER and IR can boost the

performance for the base transformer encoder, *e.g.*, $0.373 \rightarrow 0.379$ and $0.373 \rightarrow 0.382$ in BLEU-1 score respectively. As the results show, IR brings a more considerable improvement than ER (0.382 BLEU-1 vs. 0.379 BLEU-1), which indicates that exploiting the intrinsic relationship is effective when encoding radiology image regions. The performance gain of RR-Encoder comes from the similar anatomical structure of human organs, in which the extrinsic and intrinsic relationship is beneficial for extracting the inherent information of chest X-rays images.

Effect of IK-Distiller We evaluate the impact of the proposed contextual knowledge and clinical knowledge in IK-Distiller. Specifically, the distiller without contextual feature and clinical feature is a transformer-based decoder layer. Tab. 2 shows these two additional knowledge feature significantly boost the performance as a dramatic reduction can be seen when both of them are removed (from 0.393 to 0.337). In detail, the contextual feature is essential in generating long free text, such as radiology reports, since there are a certain number of tokens in a long text that is not related to the input images and should be inferred from the contexts. For the clinical feature, it helps the model to obtain a radiologist-like ability with external knowledge in generating the description of the abnormalities. With clinical features, the decoding process can be regarded as a report writing process of a radiologist, who examines the image first and employs the related clinical knowledge to complete a report. Overall, since contextual features and clinical features benefit the performance from different perspectives, adopting IK-Distiller can lead to generating more accurate reports with abnormalities descriptions.

Analysis of Connection We evaluate the role of the U-connection between the encoder and decoder layers. In the previous work, the transformer-based model has been successfully applied to the captioning tasks with the original connection structure for uni-modal scenarios like machine translation. We speculate that the generating task in cross-modal scenarios requires more specific architectures. Thus we propose U-connection and compare variations of


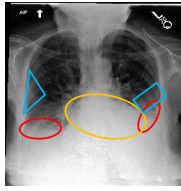
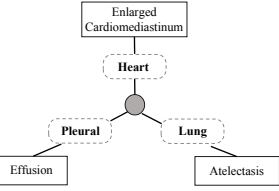

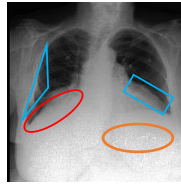
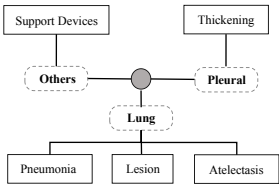
Radiology Image	Labeled Image	Related Symptom Graph	Ground Truth and Generated Report
			<p>Ground Truth There are moderate bilateral pleural effusions with overlying atelectasis, underlying consolidation not excluded. Mild prominence of the interstitial markings suggests mild pulmonary edema. The cardiac silhouette is mildly enlarged. The mediastinal contours are unremarkable. There is no evidence of pneumothorax.</p> <p>KiUT Frontal and Lateral views of the chest were obtained. There are moderate bilateral pleural effusions with overlying atelectasis underlying consolidation is not excluded. The cardiac silhouette remains enlarged. The aorta is calcified and tortuous. No overt pulmonary edema is seen.</p>
			<p>Ground Truth PA and lateral views of the chest. There is mild bibasilar atelectasis. Persistent slight elevation of the right hemidiaphragm. There is no focal parenchymal opacities concerning for pneumonia. There is no pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are stable. Multiple surgical clips in the left upper quadrant.</p> <p>KiUT Frontal and lateral views of the chest were obtained. There are low lung volumes which accentuate the bronchovascular markings. Given this there is mild elevation of the right hemidiaphragm. No definite focal consolidation is seen. There is no pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are stable. Surgical clips are seen in the upper abdomen.</p>

Figure 6. The visualization of the *KiUT*. The colored bounding regions in labeled and colored text in reports represent the abnormalities. The symptom graph represents a symptom subgraph inferred from the input image and extracted in our graph. Different colors indicate different symptoms. Different bounding shapes of one symptom mean different medical manifestations in clinical diagnosis.

the original Transformer connection structure with it. As shown in Tab. 3, we introduce different connection methods in natural image caption models and radiology report generation models. It can be seen that the 1-to-1 connection schema where the i -th decoder layer is only connected to the corresponding i -th encoder layer can bring an improvement with respect to only applying the output of the last encoder layer. Furthermore, our proposed U-connection can boost the performance of R2GenCMN (from 0.353 to 0.363 in the BLEU-1 score). More encouragingly, the U-connection schema can achieve comparable performances as the meshed connection [7] with a simple structure and fewer parameters. Thus, we confirm that exploiting the interaction between the encoder and decoder is beneficial, and the U-connection is more conformable in the cross-modal scenario than the other connection schema.

4.6. Qualitative Analysis

To better understand the effectiveness of our model, qualitative examples are given in Fig. 6. Intuitively, the reports generated by *KiUT* are accurate and robust, which show significant alignment with ground truth reports. As the figure shows, our model can learn from radiology image features and extract the relevant symptoms from the symptom graph (e.g., “enlarged cardiome-diastinum”, “pleural effusions” and “atelectasis”). It is worth noting that our model can find and describe some subtle observations, such as “support devices” and “slight elevation of the right hemidiaphragm”. This could be attributed to the well-designed U-connection for multi-level interaction and the injected knowledge. On the other hand, these examples indicate the

ability of our model to accurately discriminate and describe the different clinical manifestations of the same symptom (the blue labels). Owing to the employment of injected knowledge distiller, our model also has the certain ability to reason and write. As the example shown in Fig. 1 (*GT*: “The cardiac silhouette also appears enlarged, but it is difficult to completely assess the left border given the large pleural effusion.” and *Gen*: “The cardiac silhouette is difficult to assess due to the left base opacity.”), the sentence should be generated according to the context and the correspondence of different symptoms. More visualizations are included in the supplementary material.

5. Conclusion

In this paper, we present *KiUT*, a novel framework for radiology report generation that focuses on extracting and distilling multi-level information and multiple injected knowledge. Our model encodes images with the extrinsic and intrinsic relationships among image regions and decodes words through an injected knowledge distiller. We propose a novel U-connection schema to exploit the interaction between the encoder and decoder, which is unprecedented for other architectures in such a cross-modal scenario. Experimental results on the MIMIC-CXR and IU-Xray datasets demonstrate that our approach achieves a state-of-the-art performance and outperforms recent research from the external knowledge view. Ablation studies also prove the effectiveness of the proposed parts. We leave sophisticated knowledge constructing and structured report template filling solutions as future work.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. [2](#), [6](#)
- [2] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005. [6](#)
- [3] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022. [2](#), [5](#), [6](#), [7](#)
- [4] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. [5](#), [6](#)
- [5] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRyVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022. [4](#)
- [6] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019. [1](#)
- [7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. [1](#), [2](#), [3](#), [5](#), [7](#), [8](#)
- [8] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. [2](#), [4](#), [5](#)
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [4](#)
- [10] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020. [5](#)
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [6](#)
- [12] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32, 2019. [5](#)
- [13] Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 293–303. Springer, 2021. [2](#)
- [14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. [6](#)
- [15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Illcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. [6](#)
- [16] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1655–1663, 2021. [2](#)
- [17] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017. [1](#)
- [18] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. [2](#), [4](#), [5](#)
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. [6](#)
- [20] Changlin Li, Zhihui Li, Zongyuan Ge, and Mingjie Li. Knowledge driven temporal activity localization. *Journal of Visual Communication and Image Representation*, 64:102628, 2019. [2](#)
- [21] Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20656–20665, 2022. [1](#), [2](#)
- [22] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, pages 1–18, 2022. [2](#)
- [23] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018. [2](#)
- [24] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371, 2017. [1](#)
- [25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. [6](#)

- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [7](#)
- [27] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762, 2021. [1](#), [2](#), [5](#), [6](#)
- [28] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*, 2021. [1](#), [6](#)
- [29] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. [2](#)
- [30] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. [2](#)
- [31] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022. [2](#)
- [32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. [6](#)
- [33] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. [2](#), [6](#)
- [34] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. [1](#)
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [2](#), [4](#), [5](#)
- [36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. [1](#)
- [37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. [2](#)
- [38] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. *arXiv preprint arXiv:2207.04818*, 2022. [2](#), [5](#), [6](#), [7](#)
- [39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. [1](#), [2](#)
- [40] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466. Springer, 2018. [1](#)
- [41] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–82. Springer, 2021. [6](#)
- [42] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer, 2019. [1](#)
- [43] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15465–15474, 2021. [1](#), [2](#), [7](#)
- [44] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917, 2020. [2](#), [6](#)