Cross-modal Clinical Graph Transformer for Ophthalmic Report Generation

Mingjie Li[1,5]   Wenjia Cai[2]   Karin Verspoor[3]   Shirui Pan[5]   Xiaodan Liang[4]   Xiaojun Chang[1*]

[1]ReLER, AAII, University of Technology Sydney

[2]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University

[3]School of Computing Technologies, RMIT University

[4]School of ISE, Sun Yat-Sen University, Peng Cheng National Lab

[5]Department of Data Science and Artificial Intelligence, Monash University

Abstract

*Automatic generation of ophthalmic reports using data-driven neural networks has great potential in clinical practice. When writing a report, ophthalmologists make inferences with prior clinical knowledge. This knowledge has been neglected in prior medical report generation methods. To endow models with the capability of incorporating expert knowledge, we propose a **C**ross-modal clinical **G**raph **T**ransformer (CGT) for ophthalmic report generation (ORG), in which clinical relation triples are injected into the visual features as prior knowledge to drive the decoding procedure. However, two major common Knowledge Noise (KN) issues may affect models' effectiveness. 1) Existing general biomedical knowledge bases such as the UMLS may not align meaningfully to the specific context and language of the report, limiting their utility for knowledge injection. 2) Incorporating too much knowledge may divert the visual features from their correct meaning. To overcome these limitations, we design an automatic information extraction scheme based on natural language processing to obtain clinical entities and relations directly from in-domain training reports. Given a set of ophthalmic images, our CGT first restores a sub-graph from the clinical graph and injects the restored triples into visual features. Then visible matrix is employed during the encoding procedure to limit the impact of knowledge. Finally, reports are predicted by the encoded cross-modal features via a Transformer decoder. Extensive experiments on the large-scale FFA-IR benchmark demonstrate that the proposed CGT is able to outperform previous benchmark methods and achieve state-of-the-art performances.*

1. Introduction

Fundus Fluorescein Angiography (FFA) is one of the essential ophthalmic imaging examinations in clinical practice. However, writing reports to summarize findings from dozens of ophthalmic images during an examination is time-consuming and error-prone, especially for inexperienced ophthalmologists. With the success of data-driven neural networks [11, 19, 22, 25, 42] in many real-life scenarios, researchers and ophthalmologists start to investigate how to apply artificial intelligent (AI) models in clinical ophthalmic practice and acquire significant achievements [3]. Automatic generation of ophthalmic reports offers the possibility of reducing the heavy workloads of ophthalmologists. Furthermore, the predicted reports can highlight abnormalities for the ophthalmologists and provide a rationale for disease diagnosis; hence, automatic ophthalmic report generation has attracted increasing research interest for AI-based clinical decision support, as well as presenting a meaningful opportunity to explore the integration of vision and language modalities in neural network models.

Despite significant progress in generic image captioning models [2, 8], when transferring them into medical knowledge-driven tasks, they fail to achieve promising and competitive performance due to a lack of prior medical knowledge. When describing ophthalmic images, ordinary people can only recognize the common visual information, such as the shape and color, while ophthalmologists make inferences with their prior clinical knowledge. For models to achieve this capability, recent work explores the incorporation of medical knowledge to enhance diagnostic models [20, 23, 26, 41].

On the one hand, researchers [20, 23] have explored graph structure weights as posterior knowledge to alleviate the textual bias. In each graph, the nodes are observed abnormalities selected from prior knowledge, such as external medical corpus, and the edges are the predicted weights correlating each pair abnormalities. However, the weight graph limits the effectiveness of the knowledge graph from two aspects. Firstly, some entities are extracted from the external medical corpus or knowledge graph database separated from the training corpus. These entities will bring in a

heterogeneous embedding space [27] which makes the embedding vectors inconsistent. Secondly, there are no ground truth weights to supervise the message passing procedure, and the model is still prone to be distracted by the visual bias in medical images [26]. On the other hand, a universal graph is proposed with prior knowledge on 20 chest findings [41] to enhance models. Since these findings are not always depicted in one report, incorporating all this knowledge may divert the visual features from their original meaning.

To address these issues, we propose a **C**ross-modal clinical **G**raph **T**ransformer (CGT) for ophthalmic report generation (ORG). In particular, we first invoke an information extraction scheme based on a natural language processing pipeline, including named entity recognition and entity linking, to obtain a clinical knowledge graph. More details will be introduced in Section 3.2. As discussed in [15], the structured clinical information behind the free-text reports can enhance the diagnostic methods. In addition, the entities and relations in our clinical graph are in the homogeneous embedding space with the training corpus. Given a set of ophthalmic images, the extracted visual features are transformed to a compressed visual token and a sub-graph with relevant restored triples. Since the sub-graph is not guaranteed to be a completely accurate representation of the given images and natural noise exists in the clinical graph, we adopt a cross-modal encoder to encode the universal feature token and sub-graph information. To avoid influence from unrelated entities, a visible matrix is introduced during the cross-modal encoding process. Finally, reports are generated via a Transformer [34] decoder.

We conduct extensive experiments on the publicly available FFA-IR benchmark [21]. Experiments show that our CGT achieves the state-of-the-art performance of predicted reports under four automatic evaluation metrics and high AUC scores for the restored triples, providing a solid rationale for the explanation.

2. Related Work

2.1. Medical Report Generation

Most existing medical report generation (MRG) models are proposed to describe radiology images, especially Chest X-Ray images [17], due to the limited access to medical resources. Recently, various medical report generation datasets have been released targeting on different medical modals, such as FFA images [21], lung CT scans [23] and color fundus photography (CFP) [14], and attracted increasing attention. Inspired by work on image captioning, researchers have adopted a hierarchical recurrent network (HRNN) to describe medical images at the beginning [16,40]. In these HRNNs, the visual features extracted by a convolution neural network are attended with textual

information to generate reports. Different to generic image captioning, there are data biases stemming from both visual and textual information. The textual data bias leads to similar sentences among different reports. Therefore, Cao *et al.* [6], and Li *et al.* summarized a set of sentence templates and used the retrieved semantic features to fill the templates and generate a report. With the success of Transformer in the vision-and-language field, many Transformer-based MRG models [1, 9, 23, 30] have been proposed to replace the LSTM since Transformer is one of the most effective encoder-and-decoder frameworks. Chen *et al.* proposed a memory matrix to drive the decoding procedure. Alfarghaly *et al.* introduced 105 tags and concatenated the weighted tag embedding with visual features for decoding. Unlike others, Wang *et al.* firstly employs a selective search algorithm to extract the region-level image features to improve the MRG models. Since medical report generation is highly knowledge-driven, researchers have started incorporating medical knowledge to enhance the models.

2.2. Medical Knowledge Enhanced Models

In this section, we will introduce medical knowledge enhanced models for medical report generation and other medical domain tasks, medical QA, or memorization. The incorporated medical knowledge can be divided into three groups.

The first kind is from radiologists' working patterns [23, 26]. In clinical practice, radiologists read images and write reports in a specific pattern to remind them of not missing any part of the images. After browsing the whole image, radiologists will focus on the suspicious regions. To make the model achieve this capability, Li *et al.* introduced two kinds of auxiliary signals to guide the MRG model. Similarly, Liu *et al.* adopted both posterior and prior knowledge to imitate the pattern with retrieved reports and a universal knowledge graph. Secondly, researchers explored the clinical knowledge behind the free-text reports to drive MRG models. Both [19] and [23] extract normal and abnormal terminologies from corpus as nodes and automatically predict weights between these findings as edges to construct a knowledge graph. This graph worked as prior knowledge to drive the decoding procedure and restore a unique sub-graph for each case. In contrast, Zhang *et al.* [41] and Liu *et al.* adopted an universal graph covering 20 findings to enhance the MRG models. In the last, the existing biomedical knowledge base is adopted to incorporate medical knowledge. The unified medical language system (UMLS) [5] is the largest biomedical knowledge base and is adopted in [27] and [12] to enhance pretrained medical models for medical QA tasks. However, utilizing the existing knowledge base will bring in inconsistencies due to the heterogeneous embedding space arising from vocabulary and context mismatch. Since the entities and relations in UMLS are derived independently of the training corpus, when em-
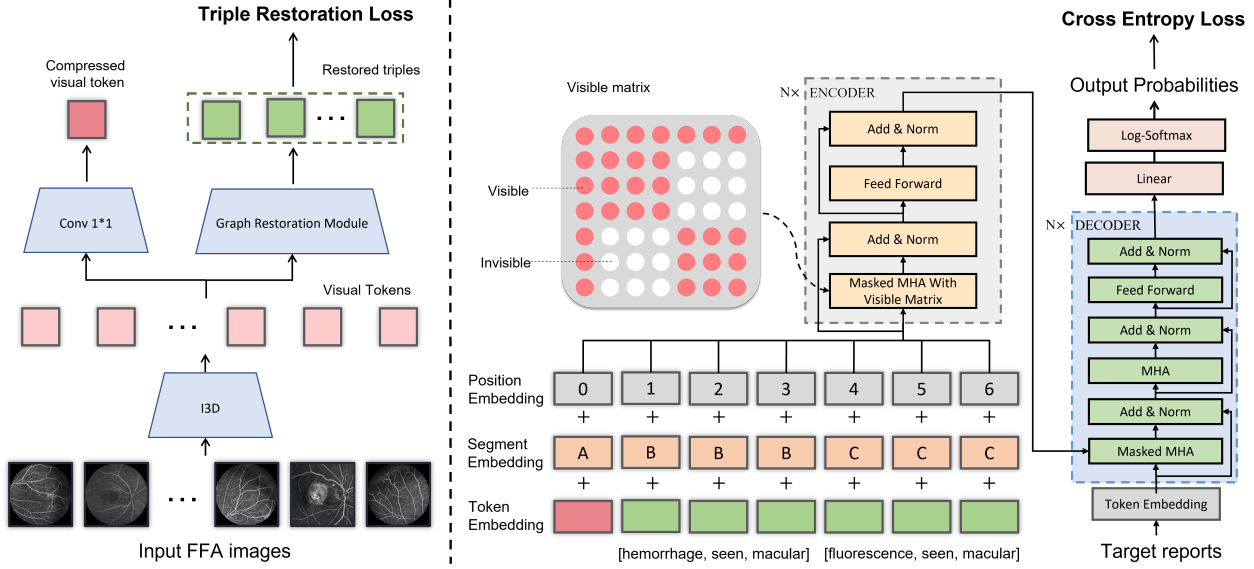
Figure 1. Illustration of our proposed cross-modal clinical graph transformer. Visual features extracted by an I3D are utilized to restore sub-graph information and compressed to one visual token. Then the cross-modal information encoded with visible matrix masked multi-head attention is used for report generation.

bedding node information, the embedded token vectors are inconsistent. Additionally, utilizing the full UMLS in MRG tasks will place a burden on the computation resources since it has $13,555,037$ triples, and most of them are irrelevant to our task.

3. Methodology

In this section, we introduce the clinical graph extraction scheme, and the process is shown in Figure 2. Then we detail the implementation of CGT (see Figure 1 for the overall framework).

3.1. Notation

In ORG task, given a set of FFA images which represented by $I = \{x_1, x_2, \ldots, x_{N_i}\}$, where x_j and N_i refer to the j-th FFA image and the number of total images, model is asked to generate a descriptive report encoded as $R = \{y_1, y_2, \ldots, y_{Nr}\}$. While we denote the ground truth report by $\hat{R} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N\hat{r}}\}$. We extract entities and relations from all the training $\hat{R}$ to construct a clinical graph (CG), denoted as $\mathcal{G}$, which is a collection of triples $\epsilon = (e_s, r, e_o)$, where e_s and e_o denote the names of subjective and objective entities, and r is the relation between them. All the triples are in CG, i.e., $\epsilon \in \mathcal{G}$. In this paper, English tokens are taken at the word-level and each token y_i, e_i and r_i are in the same vocabulary $\mathcal{V}$ whose size is d_V to make all the embedding vectors consistent.

3.2. Clinical Graph Extraction Scheme

Recently, extracting clinical information from medical reports has received increasing attention [15, 38]. The structured clinical information within the free-text reports is valuable for clinical reasoning and a variety of critical
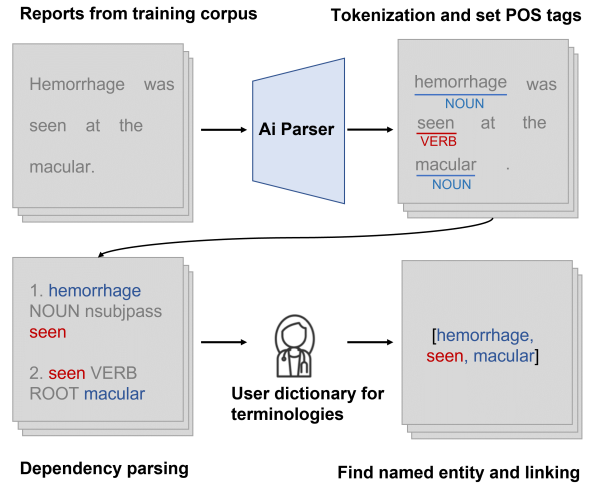

Figure 2. Process for extracting entities and relations from ophthalmic reports.

healthcare applications. We believe that ORG is one such application. However, due to the huge domain discrepancy between different medical models, transferring information from existing biomedical knowledge databases is unlikely to be effective. In this subsection, we will introduce our information extraction scheme to detail how we construct a clinical graph $\mathcal{G}$ from ophthalmic reports. This scheme is implemented by a SpaCy [29] natural language parser in an AI accelerating human-in-the-loop manner [39]. Notably, the ophthalmic reports used in this scheme are all derived from the training set to avoid target leakage.

To save the writing space, we take one sentence, "*Spotted*

Table 1. Statistics of our clinical graph.

# Entities	# Relations	# Triples
1,811	29	4,823

obscured fluorescence (hemorrhage?) was seen at the inferior edge of the macular arch ring during left eye imaging." from an ophthalmic report as an example, and the whole process is shown in Figure 2. Our scheme contains seven steps by following: **Tokenization**, taking the sentence into word-level and segmenting tokens into words, punctuation marks etc; **Part-of-speech tagging**, before automatically recognizing the relations between each pair tokens, we assign work types to each token, such as verb or nun; **Dependency parsing**, assigning syntactic dependency labels to describe the relations between individual tokens, such as '*spotted*' is the attributive of subjective '*fluorescence*'; **Lemmatization**, digging the base form of tokens. For example, the lemma of 'was' is 'is'; **Sentence boundary detection**, finding individual sentences to prevent the calculation across sentences; **Named entity recognition**, we create a user-dictionary to assist the machine in recognizing rare ophthalmic terminologies, such as 'macular'; **Entity linking**, linking entities with their relation to creating triples. Triples extracted from the sample are "*fluorescence, seen, macular*" and "*hemorrhage, seen, macular*", respectively. Then we collect all the unique triples to construct the whole clinical graph $\mathcal{G}$. In total, our clinical graph contains $4,823$ triples, and more details are presented in Table 1.

3.3. Cross-modal Clinical Graph Transformer

The traditional report generation models are based on an encoder-decoder architecture. Among all the encoder-decoder frameworks, Transformer [34] has achieved great success in various tasks. Therefore, we adopt a Transformer, the backbone of our proposed CGT, to describe ophthalmic images from the FFA-IR benchmark. As shown in Figure 1, our CGT is composed of a visual extractor, a graph construction module, a cross-modal encoder, and a language decoder.

Visual Extractor Different from describing radiology images, the average number of input images for each case is 97 in the FFA-IR. Despite the benchmark proposed by [21] is adopting lesion features via a Faster-RCNN [33], we utilize an I3D[1] model pretrained on Kinetics [7] to extract visual features from given images. Due to the reason that the entities in our CG contain both abnormalities and normal tissues, while the lesion information provided by the FFA-IR is all about the lesions or abnormalities. This data bias may mislead the message passing inter the CG.

Since the image numbers are different among each case, we first transform the given images and set a fixed length

of 96 for all the input images. For those whose length is more than 96, we randomly down-sample some images. In contrast, we repeat the whole sequence until its length is 96, when its length is below the threshold. The I3D model extracts one feature from every eight images, and the final visual features can be denoted as $f_V = \{f_1, f_2, \ldots, f_{12}\}$, where $f_i \in \mathbb{R}^{12 \times 1024}$.

Graph restoration module The graph construction module is proposed to restore a sub-graph according to the visual features generated by the visual extractor. The sub-graph encoded as $\mathcal{G}_s = \epsilon_1, \epsilon_2, \ldots, \epsilon_{Ngs}$ is a combination of triples. The whole process can be written as follows:

$$\mathcal{G}_s = max(0; BN(conv_{3 \times 3}(f_V)))W_f + b_f \qquad (1)$$

where $max(0; *)$ and BN represent the ReLU activation function and batch normalization operation, respectively; $W_f \in \mathbb{R}^{1024 \times d_V}$ denotes learnable matrix for linear transformation, while b_f refers to the bias terms. Firstly, we adopt a convolution layer with a 3×3 kernel followed by an operation sequence of batch normalization and ReLU activation to fuse the temporal information inside the f_v. Then the output has been projected by a linear transformation layer to the dimension of $d = d_V$. As mentioned, all the tokens in CG are in the same vocabulary with the training corpus; then, each vector is used to restore the index of entity or relation in $\mathcal{V}$.

Cross-modal Encoder In this module, the visual features, and the graph information are encoded by self-attention mechanism [34]. The input of the cross-modal encoder comes from the visual extractor and the graph restoration module. As mentioned in [26,37], serve visual bias exists in most medical datasets for two reasons: the abnormal regions only take a small portion of the whole image, and the human tissues are highly similar. To alleviate the impact of visual bias, we compress the f_v into one compressed visual token, encoded as $T_v \in \mathbb{R}^d$, and concatenate it with a sub-graph before fed to the embedding layer. The compressed visual token has two more advantages. Firstly, it promises that the sub-graph information is dominant to the input features. More importantly, it can be used to resist the inevitable noise inside the clinical graph adaptively since the knowledge graph can not be completely accurate.

We utilize an 'argmax' function on $\mathcal{G}_s$ and transform it into the one-hot format to represent the sub-graph, represented as $T_g = \{t_1, t_2, \ldots, t_{Nt}|t_i \in \mathbb{R}^{d_v}\}$. After concatenation, we feed the cross-modal tokens, encoded as $T = \{T_v, T_g\}$, to the embedding layer. The function of the embedding layer is to convert the cross-modal tokens into embedding representations. Similar to the BERT [10], the embedding representation of CGT is the sum of three parts. Firstly, each token in T_g is converted to an embedding vector of dimension $d = 512$ via a trainable lookup table. Different from BERT, the classification tag $[CLS]$

is replaced by T_v. Secondly, position embedding is added to the token embedding, and the formulation is written as follows:

$$PE_{pos,2i} = sin(pos/1000^{2i/d}) \qquad (2)$$

$$PE_{pos,2i+1} = cos(pos/1000^{2i/d}) \qquad (3)$$

where *pos* is the position of each token, i is the index of embedding dimension, and d is the dimension of the hidden states. Lastly, segment embedding is employed to identify each sentence. Notably, we find that most sentences in the training corpus contain two triples. Therefore, we consider every six tokens as a sentence. The T is marked with a sequence of segment tags, $\{A, B, \ldots, B, C, \ldots, C\}$, where A represents the compressed visual token.

Then the embedded tokens are encoded by a cross-modal encoder, the whole process of an encoder layer can be written as:

$$f_e(t) = BN(FFN(e_{attn}) + e_{attn}) \qquad (4)$$

$$e_{attn} = BN(MMHA(t) + t) \qquad (5)$$

Where *FFN* represents the feed forward layer, and *MMHA* represents the mask multi-head attention. The feed forward layer contains two linear layers with ReLU activation. It makes sure the dimensions of transformer input and output are the same. Another difference between our CGT and Transformer is that we adopt *MMHA* instead of *MHA* during the encoding process and introduce a visible matrix, M_v, to limit the impact of unrelated triples. The computation between unrelated triples is useless and untrue, which may also lead the changes to the original meanings. The visible matrix is presented in Figure 1, and it can limit the message passing inter the sentence or between the universal token. The *MMHA* can be written as:

$$\mathbf{h}_i^t = \text{softmax}(\frac{\mathbf{Q}_i(\mathbf{K}^t)' M_v}{\sqrt{d}})\mathbf{V}^t \qquad (6)$$

where $\{\mathbf{Q}, \mathbf{K}^*, \mathbf{V}^*\}$ are the packed d-dimensional *Query, Key, Value* vectors.

Language Decoder We adopt the vanilla Transformer decoder as our language decoder. The whole process of a decoder layer can be written as:

$$f_d(\mathbf{y}) = BN(FFN(e_{c_attn}) + e_{c_attn}) \qquad (7)$$

$$e_{c_attn} = BN(MHA(e_{attn}, f_e(\mathbf{t})) + e_{attn}) \qquad (8)$$

$$e_{attn} = BN(MMHA(\mathbf{y}) + \mathbf{y})) \qquad (9)$$

where *MMHA* represents the original masked multi-head self-attention, $\mathbf{y}$ is the input of decoder and y_t is the $t-$th input token in time step t. Cross-attention sublayer receives the output of encoder $f_e(\mathbf{t})$ and previous sublayer e_{attn}. In where, for each head, $\{\mathbf{Q}, \mathbf{K}^*, \mathbf{V}^*\}$ comes from

$\mathbf{Q} = W_q * e_{attn}$, $\mathbf{K} = W_k * f_e(\mathbf{x})$, and $\mathbf{V} = W_v * f_e(\mathbf{x})$, where W_* is the weight of a Linear layer. The $f_d(\mathbf{y})$ will be sent to a Linear & Log-Softmax layer to get the output of target sentences. Notably, only token embedding is adopted during the decoding procedure. The entire recursive generation process can be written as follows:

$$p(\hat{R}|I) = \prod_{t=1} p(\hat{y}_t|\hat{y}_1, \ldots, \hat{y}_{t-1}, I) \qquad (10)$$

Objectives Similar to the image captioning tasks, existing medical report generation approaches adopt cross-entropy loss to evaluate the differences between the predicted and the ground truth reports at the word level. Although many works attempt to explore auxiliary signals to drive the report generation, these signals can not supervise the learning process. For example, Li *et al.* [23] introduced an internal visual signal to locate the abnormal regions. However, there is no ground truth for the abnormal region bounding. Similarly, the accurate weights correlated paired findings in [19,41] are also unavailable. Therefore, the effect of auxiliary signals has been limited.

In this paper, we additionally introduce a triple restoration loss [12] to supervise the sub-graph restoration process since our clinical graph extraction scheme provides the ground truth structured information. It guarantees that the accurate graph information will be encoded with the visual features for report generation and is also what makes this method so effective. Specifically, the total loss function used in our CGT can be written as follows:

$$\mathcal{L}_{RG} = \lambda_{CE}\mathcal{L}_{CE} + \lambda_{TR}\mathcal{L}_{TR} \qquad (11)$$

where λ_{CE} and λ_{TR} are hyper-parameters balancing two terms. The first loss term $\mathcal{L}_{CE}$ is the cross-entropy loss. The second loss term is the triples restoration loss function to measure the energy of a knowledge triple. The specific process can be written as follows:

$$\mathcal{L}_{TR} = \sum_{\epsilon \in \mathcal{G}} \max(d(\epsilon) - d(f(\epsilon)) + \gamma, 0) \qquad (12)$$

where $\epsilon = (e_s, r, e_o)$, $d(\epsilon) = |e_s + r - e_o|$, $\gamma > 0$ is a margin hyper-parameter, $f(\epsilon)$ is an entity replacement operation that the subjective or objective entity in a triple is replaced and the replaced triple is an invalid triple in $\mathcal{G}$. Here, e_s, e and e_o are the indexes of the subjective, relation and objective tokens in $\mathcal{V}$.

4. Experiments

4.1. FFA-IR Benchmark

In this paper, we adopt a recently released largest ophthalmic report generation dataset to date, i.e., FFA-IR [21],

to verify the effectiveness of our approach. FFA-IR contains $10,790$ reports and $1,048,584$ FFA images. In addition, FFA-IR provides bilingual reports for each case and $12,166$ lesion bounding information, which can explain the diagnosis process. For a fair comparison, we report our results on the official splits, in which 8,016 reports/99,161 images for training, 1,069 reports/93,274 images for validation, and 1,604 reports/138,026 images for testing. Similar to the settings in Section 3.2, all the tokens are converted into lower cases, and those whose frequency of occurrence is less than three are removed, resulting in $3,241$ tokens including both words and marks. We additionally add $[PAD]$, $[SOS]$, $[EOS]$ and $[UNK]$ tags whose indexes are 0, 1, 2 and 3 into the vocabulary, resulting in $3,245$ tokens in $\mathcal{V}$. In addition, FFA-IR is the only publicly available ORG dataset based on our knowledge.

4.2. Baseline, Evaluation Metrics, and Settings

Along with the dataset, three benchmark models, which adopt spatial features via a ResNet [13], temporal features via an I3D, and lesions features via a Faster-RCNN, have been released. In this paper, we adopt the combination of I3D and Transformer as our baseline method for two reasons. On the one hand, the lesion information does not explain normal tissues, which take a considerable portion in our CG's nodes; On the other hand, the critical information inside the spatial features are easily inundated by global features during the feature fusion process [21].

We adopt two kinds of metrics to validate the effectiveness of this method. The widely used natural language generation (NLG) metrics[2], including BLEU [31], CIDEr [35], METEOR [4] and ROUGE-L [24], are adopted to evaluate the quality of predicted reports in word-level. Cider is adopted as the main metric since BLEU and METEOR are mainly used for machine translation evaluation, and ROUGE-L is designed for summaries. We also figure the micro-average of receiver operation characteristic (ROC) curve and report the area under the ROC curve (AUC) to evaluate the accuracy of the restored sub-graph.

The whole network is implemented by Pytorch [32] based on Python 3.7 and trained on two GeForce RTX 2080Ti GPUs. The images are resized to 224 before being fed into the I3D, and the batch size is 8. The maximum length of T is 90, padded with tag $[PAD]$. The embedding space for both visual and graph tokens is 512, and the dimension of the hidden states in the Transformer is also 512. Both encoder and decoder consist of six blocks and 8 heads. The ADAM [18] is utilized for optimizing all the parameters in our CGT, and the learning rate is $1e-4$. The whole network is trained for 50 epochs. We adopt greedy decoding when testing.

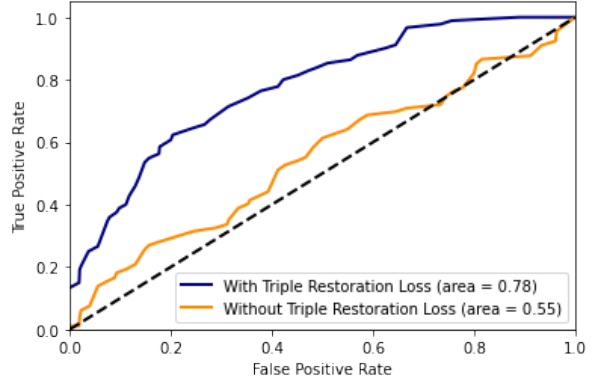[2]https://github.com/tylin/coco-caption

Figure 3. Micro-average of receiver operating characteristic curve for sub-graph restoration.

4.3. Main Results

Report generation In Table 2, we compare our CGT with a wide range of existing models. I3D+T [21] and Faster+T [21] are the two benchmark models achieving the state-of-the-art performance on FFA-IR dataset. R2Gen [9] and CoAtt [16] are the state-of-the-art radiology report generation models. The remaining presented works are from image captioning approaches. As shown in Table 2, our CGT outperforms the state-of-the-art method across all NLG metrics. The improved performance of CGT demonstrates the validity of our practice in incorporating prior medical into ophthalmic report generation.

Sub-graph restoration In Figure 3, we show the micro-average of ROC for sub-graph restoration and present the AUC scores when the proposed model is trained with triple loss restoration loss or not. With the triple restoration loss, the AUC score increased from 0.55 to 0.78 significantly. This improvement demonstrates the effectiveness of triple restoration loss and the accuracy of our restored sub-graph. Without the triple restoration loss, the restored sub-graph is similar to a sequence of random triples. It also verifies the importance of our clinical graph extraction scheme.

4.4. Quantitative Analysis

In Table 3, we present the results of quantitative analysis to investigate the contribution of each component in our CGT. The baseline model is a combination of I3D and Transformer proposed by [21].

Effect of clinical graph In this section, we evaluate the effectiveness of the proposed clinical graph, including triples and triples restoration loss.

Comparing the results in baseline and (a) in Table 3, we can find that without the triple restoration loss, the automatically restored sub-graph fails to drive the model to generate an accurate report. In (b), we randomly restore a sub-graph instead of based on the input visual features. Along with the AUC scores in Figure 3, these demonstrate that only the

Table 2. The results of NLG metrics of our proposed CGT and other state-of-the-art methods on the FFA-IR dataset. Bold numbers denote the best performance in their columns.

Methods	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
CoAtt [16]	2018	0.313	0.200	0.144	0.111	0.197	0.247	0.254
Show-Tell [36]	2015	0.306	0.197	0.142	0.109	0.191	0.247	0.232
Top-Down [2]	2018	0.320	0.217	0.162	0.127	0.207	0.289	0.363
Gounded [43]	2020	0.396	0.319	0.261	0.218	**0.229**	**0.353**	0.576
AdaAtt [28]	2017	0.292	0.181	0.127	0.095	0.205	0.236	0.234
R2Gen [9]	2020	0.330	0.225	0.167	0.132	0.210	0.296	0.367
I3D+T [21]	2021	0.428	0.341	0.276	0.229	0.213	0.334	0.561
Faster+T [21]	2021	0.443	0.355	0.288	0.240	0.205	0.341	0.590
CGT	Ours	**0.456**	**0.363**	**0.295**	**0.243**	0.227	0.345	**0.599**

Table 3. Quantitative analysis and human study of proposed method, where CVT, VM and TRL are the short for compressed visual token, visible matrix and triple restoration loss, respectively.

Settings	I3D	Triples	CVT	VM	TRL	CIDEr	BLEU-4	ROUGE	METEOR	Hit(%)
Baseline	✓					0.561	0.229	0.334	0.213	21.6
(a)		✓				0.223	0.087	0.218	0.200	-
(b)		Random				0.223	0.085	0.220	0.204	-
(c)		✓			✓	0.561	0.226	0.287	0.209	-
(d)		✓		✓	✓	0.569	0.231	0.309	**0.228**	-
(e)		✓	✓		✓	0.586	0.240	0.332	0.225	-
(f)	✓	✓			✓	0.573	0.242	0.324	0.226	-
CGT		✓	✓	✓	✓	**0.599**	**0.243**	**0.345**	0.227	**44.7**

relevant and accurate prior knowledge can improve the effectiveness of diagnostic models. Encouragingly, Table 3 Baseline and (c) illustrates that the results of utilizing the clinical graph only are competitive to the baseline. These results verify that the triples restoration loss can supervise the sub-graph restoration process and guarantee the accuracy of the incorporated prior knowledge.

Effect of visible matrix Visible matrix is another essential component in our CGT. This concept is widely used in knowledge-enhanced pretraining works [12, 27] with various formulations. In this paper, the visible matrix is adopted during the cross-modal encoding process for two purposes. On the one hand, we hope it can limit the impact of unrelated triples; On the other hand, we want the message can pass between the visual features and each triple.

The results between (c) and (d), (e) and CGT in Table 3 demonstrate the effectiveness of the visible matrix. We can see that the performances increase substantially when integrating visible matrix with (c) and (e), e.g., $0.561 \rightarrow 0.569$ and $0.586 \rightarrow 0.599$ in CIDEr score. Firstly, by comparing the results of (c) and (d), the visible matrix limits the impact from unrelated triples and greatly enhances the information interaction between related triples. Therefore, we speculate that the entity and relation representations can be well trained and improve the quality of predicted reports. When working in CGT, the visible matrix additionally facilitates

the message passing between the visual features and each triple. There is inevitable noise among the knowledge graph since the relation is not a 'hard' label. Although triple representations can be well learned, the triple may not be relevant to the input case. Therefore, the visual features play a role in de-noise adaptively. Furthermore, the visible matrix makes sure that the cross-modal signals can interact with each other.

Effect of compressed visual token The effectiveness of the compressed visual token is verified when comparing the results of (c), (e), and (f) in Table 3. As discussed, there are always noises existing in a knowledge graph. Therefore, one of the purposes for proposing a compressed visual token is to keep the accurate signals from original meanings when the sub-graph is inaccurate. When integrating the compressed visual token, the quality of predicted reports improves significantly comparing (c) and (e) and outperforming the baseline method. It demonstrates the importance of visual signals in the T. We also conducted an experiment to compare the performances of injecting prior knowledge into the compressed visual token and temporal features ((e) and (f)). We can find that the performances decrease slightly when using all the temporal features, e.g., $0.586 \rightarrow 0.573$ in the CIDEr score. We speculate the reason is that too many visual tokens will impair the effectiveness of prior knowledge. Therefore, using the compressed visual
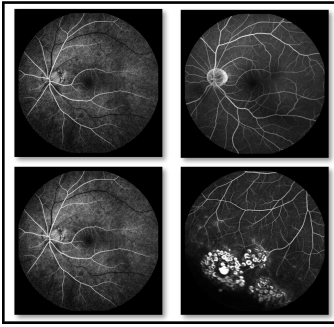
	Ground Truth:	Sub-Graph:	CGT:
	during the imaging of the left eye, some retinal microvessels in the mid-periphery of each quadrant were seen to be dilated with leakage; 2. scattered laser spot staining was seen in the inferior and mid-periphery of the temporal area during the imaging of the left eye; 3. no obvious abnormal fluorescence was seen in the retinal vessels and macular area in the posterior pole during the imaging of the left eye.	[imaging,during,eye],[fluorescence,seen,macular],[laser, spot,staining],[imaging,during,eye],[microvessel,seen,quardrant]	during the imaging of the left eye, small patches of fluorescence were seen in the macula; during the imaging of the left eye, a large number of laser spots were scattered in the retina; no significant abnormal fluorescence was seen in the retina and macular .

Figure 4. Illustrations of reports from the ground truth and CGT, and the restored sub-graph. The blue, red, and greed triples represent the true positive, false positive, and false negative.

token can make the prior knowledge dominant. Notably, the visible matrix is modified when using all temporal features. **Human study** In this section, we invited three senior ophthalmologists to evaluate the quality of predicted reports by the baseline model and our CGT. As shown in Table 3, ophthalmologists regarded that 44.7% of predicted reports by CGT can describe the given FFA images more accurately. The human study results demonstrate that our CGT outperforms the baseline model in both NLG metrics and clinical practice. Ophthalmologists also mentioned that there were 33.7% of predicted reports by both methods that failed to describe any key finding.

4.5. Qualitative Analysis

In this section, we conduct qualitative analysis for better understanding our approach via an intuitive example. Given a set of input FFA images, our CGT first restores a sub-graph which is further incorporated with visual features to generate a report.

As shown in Figure 4, one restored sub-graph consists of four triples, and each triple describes a relation between the subjective and objective entity, e.g., *[fluorescence,seen,macular]* represents that based on the prior clinical knowledge, *'fluorescence'* can be seen in the *'macular*. The number of triples is depended on the length of the input FFA images. Among the restored triples, *[fluorescence,seen,macular]* is the false positive triple which leads to the incorrect sentence *during the imaging of the left eye, small patches of fluorescence were seen in the macular.* This phenomenon shows that our CGT is capable of extending triples to a relevant sentence. Notably, due to the serve textual bias among the training corpus, the sub-graph restoration also suffers since the clinical graph is constructed from the training corpus. *[fluorescence,seen,macular]* is one of the bias triples and exists in 92% training samples. Accurately restored the triple *[laser,spot,staining]* verifies the effectiveness of our CGT to detect abnormalities among the input images and translate them into sentences. It also demonstrates that our CGT is highly capable in sub-graph restoration owing to the triple restoration loss. The last predicted sentence is not relevant to any triple in the restored sub-graph. However, this information can be provided by the compressed visual token.

5. Conclusion and Discussion

In this paper, we present an effective cross-modal clinical graph transformer for ophthalmic report generation. To obtain prior medical knowledge, we propose an information extraction scheme to construct a clinical graph from ophthalmic reports. The prior knowledge inside this graph is further restored to a sub-graph which is injected into the visual features for report generation. The experiments and analyses on the public FFA-IR dataset support our arguments and verify the effectiveness of our approach. Along with achieving state-of-the-art performances, the restored sub-graph also improves the explainability of our approach. **Negative societal impact** As with other automatic diagnostic methods, our algorithm should be utilized carefully in clinical practice since medical decisions may lead to significant consequences, including death. Therefore, while our AI diagnostic method can provide a strong rationale for judgment along with satisfactory performances, it should only be used as an auxiliary resource.

Limitations Our clinical graph is constructed in an automatic manner from a training corpus; therefore, we cannot guarantee the complete accuracy of our graph. We are inviting more experienced ophthalmologists to verify this graph. In addition, our method is not sufficiently general to support other report generation tasks. For each task, we will need to update the information extraction methods and construct a new clinical graph.

References

[1] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021. 2

[2] 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. 1, 7

[3] Maryam Badar, Muhammad Haris, and Anam Fatima. Application of deep learning for retinal image analysis: A review. *Comput. Sci. Rev.*, 35:100203, 2020. 1

[4] 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

[5] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. 2

[6] Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. Retrieve, rerank and rewrite: Soft template based neural summarization. In *ACL*, 2018. 2

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4

[8] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971, 2020. 1

[9] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020. 2, 6, 7

[10] 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

[11] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *European conference on computer vision*, pages 431–446. Springer, 2020. 1

[12] Bin He, Di Zhou, Jinghui Xiao, Qun Liu, Nicholas Jing Yuan, Tong Xu, et al. Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*, 2019. 2, 5, 7

[13] 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

[14] Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, et al. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2442–2452, 2021. 2

[15] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2, 3

[16] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. 2, 6, 7

[17] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Changlin Li, Zhihui Li, Zongyuan Ge, and Mingjie Li. Knowledge driven temporal activity localization. *Journal of Visual Communication and Image Representation*, 64:102628, 2019. 1, 2, 5

[20] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019. 1

[21] Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, et al. Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 4, 5, 6, 7

[22] Mingjie Li, Weiwei Guo, Zenghui Zhang, Wenxian Yu, and Tao Zhang. Rotated region based fully convolutional network for ship detection. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 673–676. IEEE, 2018. 1

[23] Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *arXiv preprint arXiv:2006.03744*, 2020. 1, 2, 5

[24] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, July 2004. 6

[25] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1

[26] 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, 4

[27] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020. 2, 7

[28] 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. 7

[29] Sofie Van Landeghem Matthew Honnibal, Ines Montani and Adriane Boyd. spacy: Industrialstrength natural language processing in python. 2020. 3

[30] Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*, 2021. 2

[31] 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*, July 2002. 6

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 4

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 4

[35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6

[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. 7

[37] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2433–2442, 2021. 4

[38] Joy T Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. 2021. 3

[39] Joy T Wu, Ali Syed, Hassan Ahmad, Anup Pillai, Yaniv Gur, Ashutosh Jadhav, Daniel Gruhl, Linda Kato, Mehdi Moradi, and Tanveer Syeda-Mahmood. Ai accelerated human-in-the-loop structuring of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1305. American Medical Informatics Association, 2020. 3

[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 *MICCAI*, 2018. 2

[41] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. 1, 2, 5

[42] Zenghui Zhang, Weiwei Guo, Mingjie Li, and Wenxian Yu. Gis-supervised building extraction with label noise-adaptive fully convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 17(12):2135–2139, 2020. 1

[43] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4777–4786, 2020. 7