Semi-supervised Medical Report Generation via Graph-guided Hybrid Feature Consistency

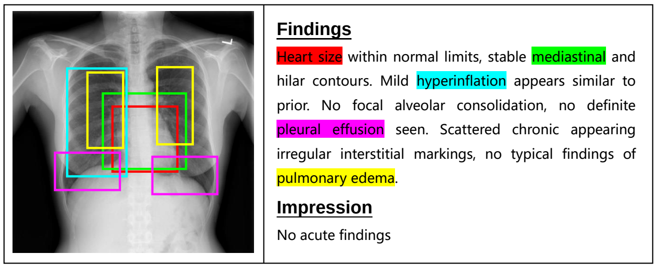
Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, *Fellow, IEEE,* Jianping Fan, Jun Yu, *Senior Member, IEEE,* and Weidong Han

***Abstract*—Medical report generation generates the corresponding report according to the given radiology image, which has been attracting increasing research interest. However, existing methods mainly adopt supervised training which rely on large amount of medical reports that are actually unavailable owing to the labor-intensive labeling process and privacy protection protocol. In the meanwhile, the intrinsic relationships between local pathological changes in the image are often ignored, which actually are important hints to high quality report generation. To this end, we propose a Relation-Aware Mean Teacher (RAMT) framework, which follows a standard mean teacher paradigm for semi-supervised report generation. The key to the encoder of the backbone network is the Graph-guided Hybrid Feature Encoding (GHFE) module, which exploits a prior disease knowledge graph to encode the intrinsic relations between pathological changes into the graph embedding and learns a word dictionary to retrieve the semantic embedding for each potential pathological change. GHFE combines the graph embedding, semantic embedding and visual features to form hybrid features, which are sent to a Transformer-based decoder for report generation. Extensive experiments on the MIMIC-CXR and IU X-Ray datasets demonstrate the effectiveness of our proposed approach.**

***Index Terms*—Medical report generation, semi-supervised learning, mean teacher, knowledge graph.**

I. Introduction

MEDICAL images and corresponding reports are widely used in disease diagnosis and treatment [1]. However, writing reports for radiologists requires strong expertise and experience, which is error-prone and becomes a heavy burden for radiologists due to the large volume of radiology images. Therefore, automatic medical report generation has gradually become a critical task in clinical practice.



Findings

Heart size within normal limits, stable mediastinal and hilar contours. Mild hyperinflation appears similar to prior. No focal alveolar consolidation, no definite pleural effusion seen. Scattered chronic appearing irregular interstitial markings, no typical findings of pulmonary edema.

Impression

No acute findings

Fig. 1. A chest X-Ray image with its report, where aligned visual regions and textual features are linked by colors.

Recently, similar to image captioning [2]–[6], several deep-learning based methods [7]–[10] for automatic report generation have been proposed and great progress has been achieved. However, there still remain two major problems in this domain.

First, due to the labor-intensive labeling process and privacy protection protocols, pairwise medical report data are hard to obtain. The scales of medical report generation datasets are relatively small, e.g., MIMIC-CXR (0.22M in capacity) [11] and IU X-Ray (4K in capacity) [12], compared with image captioning datasets, e.g., Conceptual Captions (3.3M in capacity) [13], and image pretrained datasets, e.g., ImageNet (14M in capacity) [14]. Consequently, previous works are mainly trained on very limited labeled data, which results in the heavy reliance on paired data and greatly inhibits the improvement of the model performance.

Second, the local pathological changes in a radiology image are usually not independent of one another, and they may reflect symptoms of the same type of latent disease. Therefore, there must be some kind of intrinsic relationship between these pathological changes, which are important hints to the radiologists but often ignored by current AI-based methods. In addition, for current image-report training pairs, the mapping information across radiology images and reports presented in Fig. 1 usually are not fully extracted, and the local pathological changes rarely correspond to the lexical descriptions in the report well, resulting in the inability to generate fine-grained report.

To this end, we initially propose a semi-supervised framework referring to Relation-Aware Mean Teacher (RAMT) for medical report generation with insufficient labeled data. The semi-supervised learning is realized by a standard mean teacher scheme. During the training of the student, Exponential Moving Average (EMA) is applied to the student to obtain a parameter set in temporal domain as the teacher network.

This work was supported by the National Natural Science Foundation of China under Grants 62125201, 62020106007, 62176230, 61972361. (Corresponding authors: Jun Yu and Weidong Han.)

K. Zhang and J. Yu are with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China (e-mail: ke.zhang@hdu.edu.cn; yujun@hdu.edu.cn).

H. Jiang is with Regional Medical Center for National Institute of Respiratory Diseases, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou, 310016, China (e-mail: aock@zju.edu.cn).

J. Zhang is with School of Information Science and Technology, Hangzhou Normal University, China (e-mail: jeyzhang@outlook.com).

Q. Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 101408, China (e-mail: qmhuang@ucas.ac.cn).

J. Fan is with AI Lab at Lenovo Research, 100094, China (e-mail: jfan1@Lenovo.com).

W. Han is affiliated with the Department of Medical Oncology at Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University in Hangzhou, China. Additionally, he is a professor in the College of Mathematical Medicine at Zhejiang Normal University in Jinhua, China (e-mail: hanwd@zju.edu.cn).

For any training image, we randomly perturb it to form two counterparts and feed them to the student and teacher network respectively. The student is taught to make consistent prediction with the teach at certain training iteration (when the teacher is formed). In order to learn the relations between the local pathological changes, we propose to enhance the backbone network of both student and teacher via the Graph-guided Hybrid Feature Encoding (GHFE) module with Graph Embedding (GE) and Semantic Embedding (SE). Specifically, GE interprets the local pathological changes implied by the visual features via a learned disease space and adopts a prior disease knowledge graph to encode the relations between pathological changes into graph embedding. SE interprets the semantics of the visual features via a learned medial term space, and further learns a word dictionary from which the semantic embedding for each potential pathological change can be retrieved. The graph, semantic and visual features are concatenated to form the hybrid features, which are fed to a Transformer to decode into the medical report.

It is worthwhile to highlight several aspects of this paper:

- We introduce semi-supervised learning into the field of medical report generation for the first time, and propose a novel semi-supervised framework, the Relation-Aware Mean Teacher (RAMT), to tackle this problem. RAMT reduces the reliance on large amount of labeled data and makes full use of unlabeled data.
- The GE uses disease knowledge graph to encode the relations between pathological changes into feature embedding. The SE learns a word dictionary from which the semantic embedding for each pathological change can be retrieved, such that fine-grained disease descriptions can be realized.
- We have conducted extensive experiments on two widely-used public benchmark medical report datasets, i.e., MIMIC-CXR and IU X-Ray. The results demonstrate the effectiveness of our proposed framework and enhanced backbone.

II. Related Works

A. Medical Report Generation

The automatic medical report generation works can be mainly divided into two categories: template-based retrieval methods and hierarchical encoder-decoder methods. Given an input radiology image, template-based retrieval methods retrieve the lexical descriptions about abnormalities in the image from a report database and fill the descriptions to a pre-constructed report template. In [7], the retrieval was realized by a reinforcement learning fashion with a hybrid retrieval-generation agent. KERP [8] was proposed based on abnormality graph which detected the abnormal regions in the image and retrieved the template information to explain the abnormality. Template retrieval methods not only need plenty of human resources to construct the templates but also have insufficient generalization ability because the pre-constructed template may not adapt to application scenario well.

Due to the flexibility and efficiency compared with template retrieval methods, encoder-decoder methods are much popular in recent years. TieNet [15] adopted a CNN-RNN architecture for encoding and decoding, which takes both image and text information as model input. Xue *et al.* [16] combined image encoding with generated sentences to guide the generation of the next sentence in a recurrent way. Yuan *et al.* [17] innovatively proposed a generative encoder-decoder model to synthesize multi-view visual features and incorporate medical concepts extracted from training data, which employed a synthesizing approach different from our knowledge distillation based method. A hierarchical model with Co-Attention [18] was also presented to generate reports. However, there is no annotated alignment to perform supervised training. To address this issue, reinforcement learning was introduced as a solution to guide cross-modal alignment with carefully designed rewards [19] [20].

To achieve better performance in long paragraph generation on large-scale databases, a memory-driven Transformer [9] was proposed for report generation, and it was further augmented by cross-modal memory in [21]. These methods exploit pure visual features to generate the reports. However, prior medical knowledge probably plays important role in report generation. Thus, Zhang *et al.* [10] introduced a pre-constructed graph with disease knowledge to assist report generation. Inspired by [10], PPKED [22] systematically combined the previous works to form a unified whole, including report retrieval and the use of prior knowledge. However, both of them did not specially design the network to achieve fine-grained descriptions about the pathological changes.

B. Report-supervised Representation Learning

Recently, representation learning based on medical reports has received a lot of attention and made significant progress. Among them, cross-modal pretraining is a popular research topic. Zhang *et al.* [23] pretrained medical image encoders with the paired text data via a bidirectional contrastive objective between the two modalities. Similar to [23], GLoRIA [24] used pretraining and proposed an attention-based framework for learning global and local representations by contrasting image sub-regions and words in the paired report. Zhou *et al.* [25] considered that this type of self-supervised pretraining method [23], [24] mentioned above cannot compete with the supervised paradigm. Thus, it proposed a cross-supervised methodology called REFERS that acquires free supervision signals from the original radiology reports accompanying the radiographs.

Prior work has mostly relied on the alignment of single image and report pairs even though clinical notes commonly refer to prior images. This does not only introduce poor alignment between the modalities but also a missed opportunity to exploit rich self-supervision through existing temporal content in the data. Therefore, BioViL-T [26] made use of a novel multi-image encoder and explicitly decomposed static–temporal features to augment the current image representation with information from prior images. This enables the grounding of temporal references in the text. Zhou *et al.* [27] believed that existing research has not fully utilized the complementarity between self-supervision to encode invariant semantics and associated radiology reports to incorporate medical expertise. MRM was proposed to formalize the radiograph understanding and radiology report comprehension as two complementary masked modeling objectives following a multi-task scheme to learn knowledge-enhanced semantic representations.

Overall, different from these ideas that use a two-stage strategy of pre-training and fine-tuning on downstream tasks, our method adopts an end-to-end training strategy of supervised and unsupervised tasks simultaneously. Our method does not require additional pre-training and achieves state-of-the-art performance in medical report generation.

C. Consistency-based Semi-supervised Learning

So far as we know, there is still no semi-supervised medical report generation methods. Nevertheless, semi-supervised medical image analysis has been a research hotspot recently [28], [29]. Current methods includes: adversarial-based methods [30], graph-based methods [31] and consistency-based method [32]. The adversarial-based methods generate samples using GAN and predict pseudo labels for these samples, which are used to further enlarge the training set for model update. The unsatisfactory pseudo labels will impede the report generation. Graph-based methods use graph to propagate the labels to unlabeled data. However, maintaining the whole graph requires much storage resources, which is particularly notable in case of large data amounts.

In comparison, consistency-based methods avoid the aforementioned problems. These methods perturb the unlabeled data to obtain two visually different but semantically consistent counterparts, and force the network(s) to produce consistent predictions for the two counterparts. The Temporal Ensembling (TE) method [33] directly makes the outputs of the model to be consistent with each other, and uses exponential moving average (EMA) to integrate the model output about the unlabeled data in temporal domain to obtain more confident predictions that can be treated as supervision information. However, due to the requirement of maintaining the historical data during training, it can be extremely slow for model training on large-scale datasets. To address this problem, the Mean Teacher (MT) framework [34] applies EMA to ensemble historical network parameters to define teacher networks. With the parameter ensembling, the teacher networks can also provide reliable predictions as supervision signals, but greatly reduce the cost of data maintenance during training. Inspired by [34], Liu *et al.* [35] further captures the intrinsic relevance of visual features as additional supervision information.

III. Method

According to the above-mentioned discussion, we propose a novel semi-supervised framework for medical report generation in case of insufficient training reports named as the Relation-Aware Mean Teacher (RAMT). RAMT follows the basic concern of mean teacher method to take full advantage of the unlabeled training radiology images. The backbone network is enhanced by the Graph-guided Hybrid Feature Encoding (GHFE) module, which includes Graph Embedding (GE), Semantic Embedding (SE) and feature combination. The hybrid features are sent to a Transformer-based decoder to generate the reports. This framework can be illustrated by Fig. 2. The main notations and definitions are shown in Table AI for better readability in the Appendix.

A. Semi-supervised Report Generation

We denote the labeled dataset as $D_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, and the unlabeled dataset as $D_U = \{\mathbf{x}_i\}_{i=N+1}^{N+M}$, where $\mathbf{x}_i$ is the input radiology image and $\mathbf{y}_i$ represents its corresponding medical report. According to the principle of mean teacher approach, we conduct Exponential Moving Average (EMA) to ensemble the parameters of the student network at each iteration during training to obtain a sequence of parameters, which represents certain network state with temporal contextual information and we regard the network with corresponding parameters as a teacher. For any noise perturbed $\mathbf{x}_i$ in both D_L and D_U, we expect the network to generate consistent results with the teacher at specific iteration when the teacher is formed. Thus, the current network being optimized actually plays the role of a student. Obviously, the student and teachers share the same network structure but have distinct parameters. We denote the network structure as F and characterize the student and teacher with distinct parameter set θ_s and θ_t.

The parameter ensembling can be described as:

$$\theta_t^k = \alpha \theta_t^{k-1} + (1-\alpha)\theta_s^k, \tag{1}$$

where α is a hyperparameter used to smooth the update and k means the training iteration.

The supervised loss is generated via the cross entropy between the predicted reports $\mathbf{p}_i = \mathrm{F}_{\theta_s}(\mathbf{x}_i)$ by the student and the ground truth reports $\mathbf{y}_i$:

$$\mathcal{L}_S = -\sum_{i=1}^{N}\sum_{j=1}^{L}\sum_{c=1}^{C} \mathbf{y}_{ij}(c)\log\left(\mathbf{p}_{ij}(c)\right), \tag{2}$$

where $\mathbf{y}_{ij}$ is the j-th word vector of i-th report, $\mathbf{p}_{ij}$ is the counterpart in the predicted report, C denotes the vocabulary size.

The unsupervised loss $\mathcal{L}_U$ in this paper consists of two parts, which are the report consistency loss and the intermediate consistency loss corresponding to the yellow and red dashed lines in Fig. 2. The first part is the report consistency loss $\mathcal{L}_R$, which means that $\mathbf{p}_i = \mathrm{F}_{\theta_s}\left(\mathbf{x}_i^{\tau_s}\right)$ should be consistent with the predicted reports $\mathbf{q}_i = \mathrm{F}_{\theta_t}\left(\mathbf{x}_i^{\tau_t}\right)$ by the teacher at each iteration, where $\mathbf{x}_i^{\tau_s}$ and $\mathbf{x}_i^{\tau_t}$ represents the perturbed $\mathbf{x}_i$ by noise τ_s and τ_t respectively. Consequently, we formulate $\mathcal{L}_R$ as the Kullback–Leibler divergence between $\mathbf{p}_i$ and $\mathbf{q}_i$:

$$\mathcal{L}_R = \sum_{k=1}^{T}\sum_{i=1}^{N+M}\sum_{j=1}^{L}\sum_{c=1}^{C} \mathbf{q}_{ij}^{k}(c)\log\left(\frac{\mathbf{q}_{ij}^{k}(c)}{\mathbf{p}_{ij}^{k}(c)}\right). \tag{3}$$

In addition to report consistency, we also require the intermediate layer output of the student is consistent with that of the teacher at each iteration, and this is referred to as the intermediate consistency loss $\mathcal{L}_I$. Let $\mathbf{h}_s = \tilde{\mathrm{F}}_s\left(\mathbf{x}_i^{\tau_s}\right)$ and $\mathbf{h}_t = \tilde{\mathrm{F}}_t\left(\mathbf{x}_i^{\tau_t}\right)$ be the vectorized intermediate representation of the student and teacher, $\tilde{\mathrm{F}}_s$ and $\tilde{\mathrm{F}}_t$ are the subnetworks including GHFE module to generate the hybrid features, which

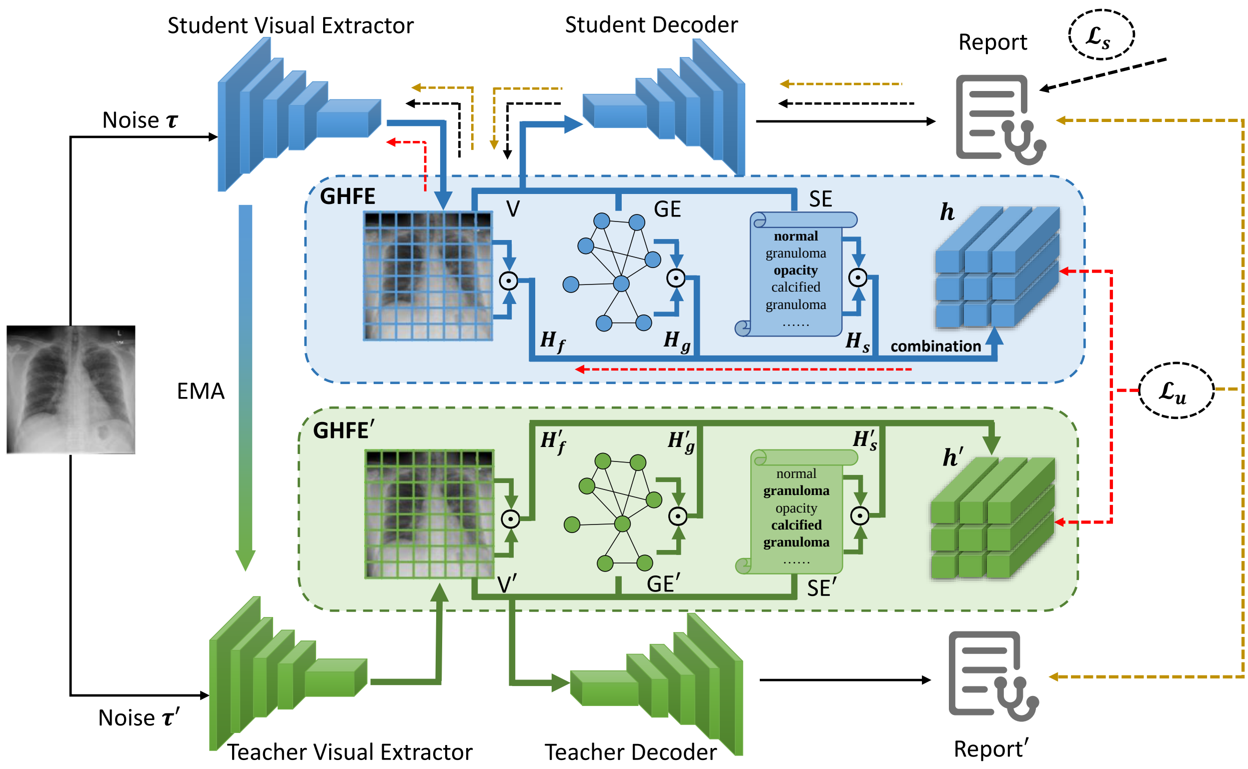


Fig. 2. The architecture of the Relation-Aware Mean Teacher (RAMT) framework. Exponential Moving Average (EMA) is applied to student to obtain teacher network that guides the student to be consistent with it for noise (τ and τ') perturbed samples. GHFE/GHFE' generates the hybrid features by concatenating GE, SE and visual features V, which are fed to decoders for report generation. The unsupervised consistency constraints $\mathcal{L}_u$ are applied to both the predicted reports and hybrid features. The supervised constraints $\mathcal{L}_s$ rely on the ground truth reports. The dashed lines represent the backpropagation of gradients.

will be further discussed in next section. $\mathcal{L}_I$ can be denoted as:

$$\mathcal{L}_I = \sum_{k=1}^{T} \sum_{i=1}^{N+M} \left\| (\mathbf{h}_s)_i^k - (\mathbf{h}_t)_i^k \right\|_2 . \tag{4}$$

The overall objective function of this framework is defined as:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_U = \mathcal{L}_S + \lambda(\mathcal{L}_R + \beta \mathcal{L}_I), \tag{5}$$

where λ and β are the trade-off weights to balance the ratio of different losses.

B. Graph-guided Hybrid Feature Encoding

Inspired by [9], we use the pure Encoder (CNN)-Decoder (Transformer) architecture as the basic backbone network for both the student and the teachers, as shown in Fig. 3(a). In order to verify the effectiveness of the proposed methods later, we remove all the memory-related modules. The key to the Encoder part is the Graph-guided Hybrid Feature Encoding (GHFE) module, which consists of Graph Embedding (GE), Semantic Embedding (SE) and feature combination, refer to Fig. 3(b). Since the Encoder of the student and teachers follows the exactly same working pipeline, we will introduce the detailed description without distinction.

1) Graph Embedding: It is commonly accepted that multiple pathological changes in a radiology image are not completely independent, and they may jointly imply the appearance of the same disease. Thus, the relationships between the pathological changes are important information to report generation. Detecting these pathological changes needs additional training and deduction, which degrades the computational efficiency and accuracy. Accordingly, we follow [10] to use a pre-constructed disease knowledge graph to learn relationships between the pathological changes, and this module is named as Graph Embedding module (GE).

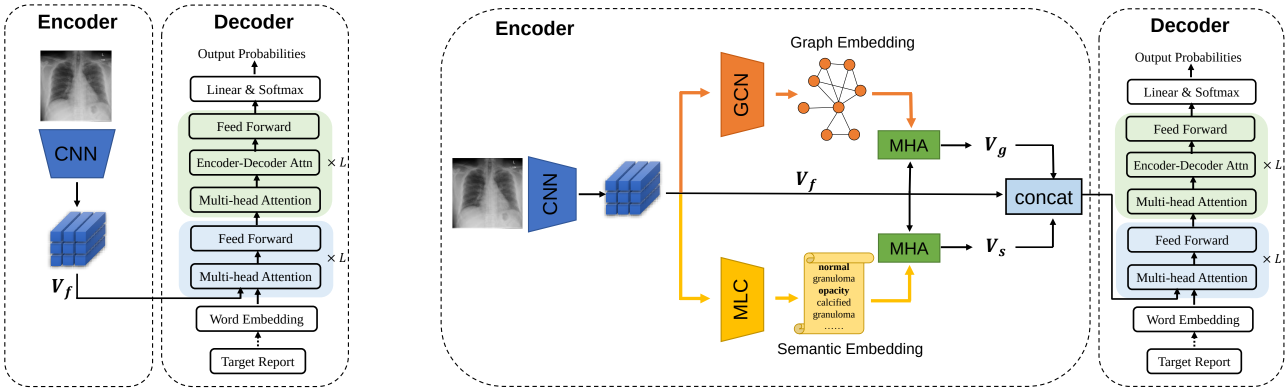
GE first projects the visual features $\mathbf{V}_f \in \mathbb{R}^{c \times H \times W}$ of an image extracted by pre-trained DenseNet121 [36] into a learnable disease space $\mathbf{Q} \in \mathbb{R}^{d \times c}$ ($\mathbf{Q}$ can be randomly initialized), where d represents the number of disease categories in the pre-defined graph and it is set to 20 following [10]. This is to interpret the pathological changes implied in $\mathbf{V}_f$ with potential diseases. Then, GE applies softmax to obtain the probabilities $\mathbf{E}$ that these diseases occur at each position of the feature maps:

$$\mathbf{E} = \mathrm{softmax}\left(\mathbf{Q} \cdot \mathbf{V}_f\right). \tag{6}$$

The visual features are then enhanced by attention as $\mathbf{F} = \mathbf{E} \cdot \mathbf{V}_f^T \in \mathbb{R}^{d \times c}$. We stack the average features for each channel of $\mathbf{V}_f$ to $\mathbf{F}$ to initialize the graph node features:

$$\mathbf{G}^0 = \left[\frac{1}{H \times W} \sum_{i,j} \left(\mathbf{V}_f^T\right)_{i,j,c}, \mathbf{F} \right] \in \mathbb{R}^{(d+1) \times c}. \tag{7}$$

Let $\tilde{\mathbf{A}} = \left[\mathbf{I}, \mathbf{A}, \mathbf{A}^T\right]^T$ where $\mathbf{A} \in \mathbb{R}^{(d+1)\times(d+1)}$ is pre-constructed disease knowledge Laplacian matrix that defines the relationships between typical prior diseases, $\mathbf{W}^l =$



(a) Basic backbone network

(b) Enhanced backbone network

Fig. 3. The basic backbone network and the enhanced backbone with Graph-guided Hybrid Feature Encoding (GHFE).

$\left[\mathbf{W}_1^l,\mathbf{W}_2^l,\mathbf{W}_3^l\right]\in\mathbb{R}^{c\times 3s}$, $\left(\mathbf{W}_i^l\in\mathbb{R}^{c\times s}\right)$ are parameters, $\mathbf{G}^0$ can be subsequently updated through graph convolutional networks:

$$\left\{\begin{array}{c}\tilde{\mathbf{G}}=\operatorname{Diag}\left(\operatorname{ReLU}\left(\operatorname{Norm}\left(\tilde{\mathbf{A}}\cdot\mathbf{G}^l\cdot\mathbf{W}^l\right)\right)\right)^T\\ \mathbf{G}^{l+1}=\operatorname{ReLU}\left(\operatorname{Norm}\left(\tilde{\mathbf{G}}\cdot\mathbf{W}^{l+1}\right)\right)\end{array}\right., \tag{8}$$

where $\operatorname{Diag}(\cdot)$ extracts 3 diagonal subblocks of size $(d+1)\times s$ such that $\tilde{\mathbf{G}}$ is a $(d+1)\times 3s$ matrix, $\mathbf{W}^{l+1}\in\mathbb{R}^{3s\times c}$ are also parameters.

The Graph Embedding $\mathbf{V}_g\in\mathbb{R}^{c\times H\times W}$ is constructed by enhancing $\mathbf{V}_f$ through $\mathbf{G}^L$ that encodes the relationships between different pathological changes:

$$\mathbf{V}_g=\operatorname{FFN}(\operatorname{MHA}(\mathbf{V}_f^T,\mathbf{G}^L)), \tag{9}$$

where FFN means feed-forward network and MHA represents multi-head attention.

2) Semantic Embedding: In order to further express the semantic implied by the visual features $\mathbf{V}_f$, we interpret the semantic of $\mathbf{V}_f$ using a learned medical term space $\mathbf{J}\in\mathbb{R}^{d\times c}$, which is pre-trained on the MIMIC dataset for the multi-label classification (MLC) task to predict medical tags. Then $\operatorname{softmax}$ was used to obtain the probability vector $\mathbf{o}$ that indicates the chance each term occurs:

$$\mathbf{o}=\operatorname{softmax}\left(\mathbf{J}\cdot\overline{\mathbf{V}}_f\right), \tag{10}$$

where $\overline{\mathbf{V}}_f=\frac{1}{H\times W}\sum_{i,j}\left(\mathbf{V}_f\right)_{c,i,j}\in\mathbb{R}^c$. Let $\mathbf{o}_i$ be an element in $\mathbf{o}$, we transform $\mathbf{o}$ to $\tilde{\mathbf{o}}$ by setting $\mathbf{o}_i=1$ if $\mathbf{o}_i$ belongs to the top-K leading elements, and setting $\mathbf{o}_i=0$ if not.

Suppose $\mathbf{D}\in\mathbb{R}^{d\times c}$ is a learnable embedding operation mapping the dimensions of the medical tags predicted by $\mathbf{J}$ (i.e., the vocabulary size in the embedding) back to the semantic feature dimensions (i.e., the embedding size). We derive the semantic features $\mathbf{S}$ by retrieving $\mathbf{D}$ with $\tilde{\mathbf{o}}$:

$$\mathbf{S}=\operatorname{diag}(\tilde{\mathbf{o}})\cdot\mathbf{D}, \tag{11}$$

where $\operatorname{diag}(\cdot)$ expands a vector to the diagonal elements of a diagonal matrix. That is, retrieving the corresponding embeddings from the vocabulary for the top-K positions. We simplify $\mathbf{S}$ to $\tilde{\mathbf{S}}\in\mathbb{R}^{K\times c}$ by removing the zero lines, then the Semantic Embedding $\mathbf{V}_s\in\mathbb{R}^{c\times H\times W}$ can be similarly denoted as:

$$\mathbf{V}_s=\operatorname{FFN}(\operatorname{MHA}(\mathbf{V}_f^T,\tilde{\mathbf{S}})). \tag{12}$$

Both $\mathbf{Q}$ and $\mathbf{D}$ can be randomly initialized and learned in an end-to-end fashion.

3) Feature Combination: To encode both the graph, semantic and visual information into feature representation, we concatenate the graph embedding, semantic embedding and visual features to obtain the hybrid feature representation $\mathbf{V}_{hybrid}=[\mathbf{V}_f,\mathbf{V}_g,\mathbf{V}_s]$, which will be sent to the decoder for report generation.

To further represent the intermediate layer output consistency in Fig. 2, we stack $\mathbf{V}_{f/g/s}$ column by column to reshape it to $\tilde{\mathbf{V}}_{f/g/s}\in\mathbb{R}^{(c\times H\times W)\times 1}$ such that the features of a batch of data can be denoted as $\tilde{\mathbf{V}}_{f/g/s}^B\in\mathbb{R}^{(c\times H\times W)\times B}$. We deem gram matrix as $\mathbf{H}_{f/g/s}=(\tilde{\mathbf{V}}_{f/g/s}^B)^T\cdot\tilde{\mathbf{V}}_{f/g/s}^B$ to reflect the sample correlations. Let $\mathbf{H}=[\mathbf{H}_f,\mathbf{H}_g,\mathbf{H}_s]^T$, $\mathbf{H}$ can be rewritten as $\mathbf{H}=\left[\frac{\mathbf{h}_1}{\|\mathbf{h}_1\|_2},\ldots,\frac{\mathbf{h}_B}{\|\mathbf{h}_B\|_2}\right]$ where $\mathbf{h}_i$ is a column vector of $\mathbf{H}$ corresponding to a training sample and $\frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2}$ is the feature correlation representation. We reuse $\mathbf{h}_i$ to represent $\frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2}$ for simplicity in previously discussed $\mathcal{L}_I$.

4) Transformer-based Decoder: We use standard Transformer as the report generator to receive hybrid feature representation $\mathbf{V}_{hybrid}$ as input. The prediction process of the report can be denoted as follows:

$$\mathbf{y}_{ij}=F_{\theta_D}^D\left(F_{\theta_E}^E\left(\mathbf{V}_{hybrid}\right),\mathbf{y}_{i1},\mathbf{y}_{i2},\ldots,\mathbf{y}_{ij-1},\mathbf{V}_{hybrid}\right), \tag{13}$$

where $F_{\theta_E}^E$ and $F_{\theta_D}^D$ represent the encoder and decoder part of the Transformer with parameters θ_E and θ_D respectively, and $\mathbf{y}_{ij}$ is the j-th word of the i-th report. For each report $\mathbf{y}_i$, the optimal parameters $\hat{\theta}$ of the report generator maximize following conditional log-likelihood:

$$\hat{\theta}=\underset{\theta}{argmax}\sum_{j=1}^{L}\log\mathbf{p}(y_{ij}|y_{i1},y_{i2},\ldots,y_{ij-1},\mathbf{V}_{hybrid}), \tag{14}$$

where $\theta=\{\theta_D,\theta_E\}$ and L denotes the length of the report.

IV. Experiments and Results

We conduct a group of comparative and ablation experiments on two widely-used benchmark datasets i.e., IU X-Ray [12] and MIMIC-CXR [11] to evaluate the effect of the proposed method. This section firstly introduces the datasets, then the evaluation metrics and experimental setting with compared methods, followed by the results and discussion about the comparative and ablation experiments. Qualitative results with more case studies and sensitivity analysis are shown in the last to give intuitive insight to readers.

A. Dataset

- **IU X-Ray:** Indiana University Chest X-ray Collection is a public dataset to evaluate the performance of report generation methods. It contains 7,470 chest x-ray images and 3,955 corresponding reports. Following previous works [7], [8], we exclude the images without reports and randomly split the pairwise data by the ratio of 7:2:1 into training, validation and testing set.
- **MIMIC-CXR:** the largest radiology dataset which contains 373,057 chest x-ray images and 206,563 reports from 63,478 patients. Following the official paradigm [9], the dataset is split into training set with 368,960 images and 222,758 reports, validation set with 2,991 images and 1,808 reports, and testing set with 5,159 images and 3,269 reports.

The detailed characteristics of the two datasets are shown in Table AII in the Appendix.

B. Evaluation Metrics

We adopt the widely-used evaluation toolkit [37] to calculate the matching degree between generated reports and ground truth reports, which is reflected by the NLG metrics including BLEU [38], METEOR [39] and ROUGE-L [40]. BLEU and METEOR are originally designed for machine translation while ROUGE-L is originally proposed to evaluate text summarization. Both machine translation and text summarization focus more on sentence-level matching, and each word has the same contribution. However, the key to a medical report is usually the descriptions of some pathological changes that appear in the images. Accordingly, we also adopt Clinical Efficacy (CE) metrics that directly compare the phrases describing the pathological changes of the generated and ground truth report [9].

C. Experimental Settings

We adopt DenseNet121 [36] pretrained on ImageNet [14] as the visual feature extractor with dimension 1024. The report generator is a randomly initialized standard Transformer. We use Adam optimizer with batch size of 64, consisting of 16 labeled pairs (25%) and 48 unlabeled images (75%) for semi-supervised training. Following [35], the perturbance τ is realized by applying random rotation, translation, horizontal flips to images, and the drop rate of the dropout layer for visual extractor is 0.2. The learning rate is set to 5e-5 and 1e-4 for the visual extractor and other parameters respectively, which decays by a factor of 0.8 per epoch. We set the ramp-up epoch as 10 and 30 for IU X-Ray and MIMIC-CXR, respectively. λ is controlled by a Gaussian warming up function as $\lambda(t) = 1 * e^{-5(1-t/T)^2}$, which will ramp-up from 0 to 1 in first T epochs to avoid the domination of unsupervised training since its initial targets are not reliable enough.

D. Compared Methods

In this section, we briefly introduce several state-of-the-art methods compared in the next experiments.

1) Top-Down [41]: It presents a novel combined bottom-up and top-down visual attention mechanism which enables attention to be calculated more naturally at the level of objects and other salient regions.

2) HRGR-Agent [7]: It introduces a novel Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) to perform robust medical image report generation. It is also the first attempt to bridge human prior knowledge and generative neural network via reinforcement learning.

3) CMAS-RL [19]: CMAS-RL aims to exploit the structure information in the reports. Specifically, it explicitly models the between-section structure by a two-stage framework and implicitly captured the within-section structure with a Cooperative Multi-Agent System (CMAS) in a reinforced manner.

4) R2Gen [9]: R2Gen applies memory-driven Transformer to generate radiology reports, where a relational memory is used to record the information from previous generation processes and a layer normalization mechanism is designed to incorporate the memory into Transformer.

5) R2GenCMN [21]: Upgraded version of R2Gen. It focuses more on cross-modal mappings rather than the aspect of text generation, and a cross-modal memory network (CMN) is proposed to enhance the encoder-decoder paradigm, where a shared memory is designed to record the alignment between images and texts so as to facilitate the interaction and generation across modalities.

6) PPKED [22]: It presents to explore and distill posterior and prior knowledge for radiology report generation which imitates the working patterns of radiologists to alleviate the data bias problem. Due to multi-task learning with multi-branch modules, the overall model is relatively complex.

7) CMM+RL [20]: It introduces an approach with reinforcement learning (RL) over a cross-modal memory (CMM) to better align visual and textual features. In detail, a shared memory is used to store the cross-modal information and RL is applied to guide cross-modal mappings so as to better link features from images and texts.

E. Overall Results and Analyses

The comparison results on IU X-Ray and MIMIC-CXR are reported in Table I and Table II respectively, where Top-Down [41], HRGR-Agent [7], CMAS-RL [19], R2Gen [9], R2GenCMN [21], PPKED [22] and CMM+RL [20] are fully supervised methods trained with all samples each with a ground truth report, BASE+GHFE is the improved backbone depicted by Fig. 3(b) and trained with 25% samples each with a report, RAMT is the proposed semi-supervised method

TABLE I
COMPARATIVE EXPERIMENTS ON IU X-RAY DATASET, WHERE THE RESULTS ABOUT NLG METRICS OF THE PROPOSED RAMT AND OTHER STATE-OF-THE-ART METHODS WITH DIFFERENT RATIOS OF LABELED TRAINING SAMPLES ARE SHOWN. BASE REPRESENTS THE BASIC BACKBONE DEPICTED BY FIG. 3(A) AND BASE+GHFE IS THE IMPROVED BACKBONE DEPICTED BY FIG. 3(B). L, U, B-N, M AND R-L DENOTE LABELED DATA, UNLABELED DATA, BLEU-N, METEOR AND ROUGE-L, RESPECTIVELY.

MODEL	RATIO(%)		NLG METRICS					
	L	U	B-1	B-2	B-3	B-4	M	R- L
HRGR-Agent [7]	100	0	0.438	0.298	0.208	0.151	-	0.322
CMAS-RL [19]	100	0	0.464	0.301	0.210	0.154	-	0.362
SentSAT+KG [10]	100	0	0.441	0.291	0.203	0.147	-	0.367
R2Gen [9]	100	0	0.470	0.304	0.219	0.165	0.187	0.371
R2GenCMN [21]	100	0	0.475	0.309	0.222	0.170	0.191	0.375
PPKED [22]	100	0	**0.483**	**0.315**	**0.224**	**0.168**	-	0.376
CMM+RL [20]	100	0	**0.494**	**0.321**	**0.235**	**0.181**	**0.201**	**0.384**
BASE + GHFE	25	0	0.433±0.007	0.278±0.004	0.197±0.004	0.148±0.003	0.176±0.002	0.354±0.004
RAMT	25	75	0.480±0.005	0.302±0.005	0.214±0.003	0.159±0.003	**0.196**±0.001	0.368±0.006
RAMT-Upperbound	100	0	0.482±0.009	0.310±0.006	0.221±0.004	0.165±0.003	0.195±0.002	**0.377**±0.003

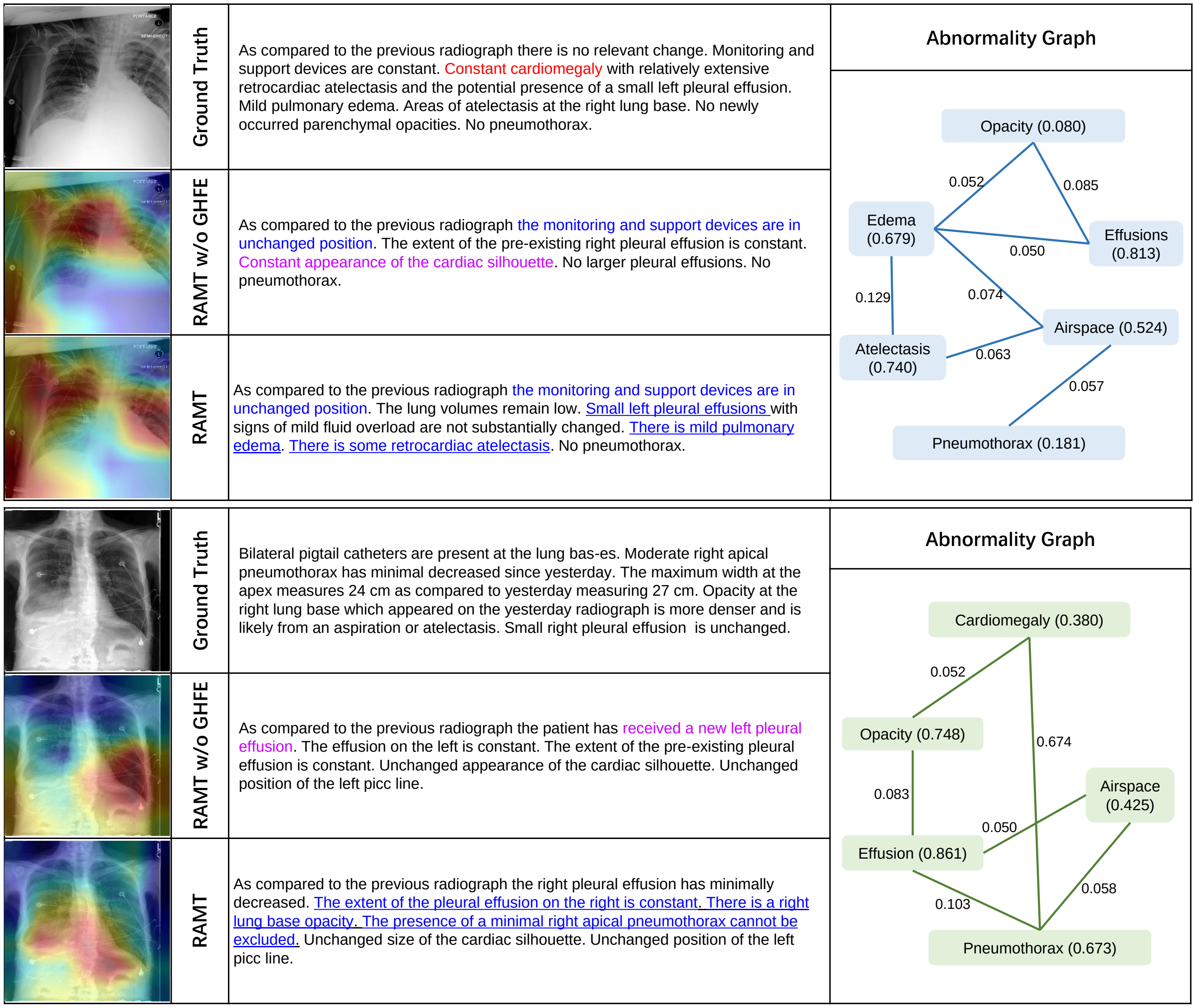
TABLE II
COMPARATIVE EXPERIMENTS ON MIMIC-CXR DATASET, WHERE THE RESULTS ABOUT NLG AND CE METRICS OF THE PROPOSED RAMT AND OTHER STATE-OF-THE-ART METHODS WITH DIFFERENT RATIOS OF LABELED TRAINING SAMPLES ARE SHOWN. P, R AND F1 DENOTE PRECISION, RECALL AND F1 SCORE, RESPECTIVELY.

MODEL	RATIO(%)		NLG METRICS						CE METRICS		
	L	U	B-1	B-2	B-3	B-4	M	R-L	P	R	F1
Top-Down [41]	100	0	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
R2Gen [9]	100	0	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
R2GenCMN [21]	100	0	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
PPKED [22]	100	0	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
CMM+RL [20]	100	0	**0.381**	**0.232**	**0.155**	**0.109**	0.151	**0.287**	0.342	0.294	0.292
BASE + GHFE	25	0	0.323±0.008	0.202±0.006	0.136±0.005	0.097±0.003	0.134±0.002	0.278±0.003	0.341±0.007	0.276±0.005	0.264±0.005
RAMT	25	75	0.358±0.005	0.221±0.004	0.148±0.002	0.106±0.003	**0.153**±0.001	**0.289**±0.003	**0.362**±0.005	**0.304**±0.006	**0.309**±0.002
RAMT-Upperbound	100	0	**0.362**±0.007	**0.229**±0.006	**0.157**±0.003	**0.113**±0.002	**0.153**±0.004	0.284±0.004	**0.380**±0.008	**0.342**±0.006	**0.335**±0.007

TABLE III
ABLATION STUDY ABOUT OUR GHFE MODULE APPLIED TO SEMI-SUPERVISED TRAINING (RAMT) AND FULLY SUPERVISED TRAINING ON MIMIC-CXR DATASET.

MODE	MODEL	B-1	B-2	B-3	B-4	M	R-L	P	R	F1
Semi-Supervised (RAMT)	V (Visual)	0.335	0.207	0.140	0.097	0.139	0.276	0.331	0.266	0.262
	V + GE	0.354	0.218	**0.149**	0.102	0.144	0.280	0.356	0.291	0.286
	V + SE	0.352	0.220	0.148	0.101	0.148	0.286	0.352	0.283	0.272
	GHFE	**0.358**	**0.221**	0.148	**0.106**	**0.153**	**0.289**	**0.362**	**0.304**	**0.309**
Fully-Supervised	V (Visual)	0.348	0.216	0.144	0.101	0.144	0.278	0.347	0.281	0.268
	V + GE	0.355	0.223	0.152	0.108	0.150	0.279	0.363	0.303	0.297
	V + SE	0.358	0.225	0.151	0.110	**0.156**	**0.284**	0.369	0.317	0.299
	GHFE	**0.362**	**0.229**	**0.157**	**0.113**	0.153	**0.284**	**0.380**	**0.342**	**0.335**

trained with 25% labeled samples and 75% unlabeled samples. RAMT-Upperbound is equivalent to training the BASE+GHFE using all labeled samples, therefore it is supposed to have the best performance among three designed methods. In addition to NLG metrics, we also applied CE metrics following [9], which adopts CheXpert [42] to label the 14 predefined diseases and calculate the Precision, Recall and F1 scores on MIMIC-CXR. For [41], we cite the result from [9]. As for other methods which are specifically designed for radiology report generation, we directly report the results from the original papers. Owing to the semi-supervised learning ability, it is predictable that the RAMT framework outperforms the BASE+GHFE under all metrics by a large margin on both datasets. More importantly, on both datasets, the RAMT yields similar results to the state-of-the-art methods in NLG metrics and outperforms the complex PPKED and CMM+RL in CE metrics, which demonstrates the effectiveness of RAMT on combining consistency mechanism and hybrid features for report generation. In addition, RAMT even yields better results than fully supervised R2GenCMN which is the augmented version of R2Gen. It is surprising that the RAMT yields even better results than the RAMT-Upperbound in METEOR and



As compared to the previous radiograph there is no relevant change. Monitoring and support devices are constant. Constant cardiomegaly with relatively extensive retrocardiac atelectasis and the potential presence of a small left pleural effusion. Mild pulmonary edema. Areas of atelectasis at the right lung base. No newly occurred parenchymal opacities. No pneumothorax.

As compared to the previous radiograph the monitoring and support devices are in unchanged position. The extent of the pre-existing right pleural effusion is constant. Constant appearance of the cardiac silhouette. No larger pleural effusions. No pneumothorax.

As compared to the previous radiograph the monitoring and support devices are in unchanged position. The lung volumes remain low. Small left pleural effusions with signs of mild fluid overload are not substantially changed. There is mild pulmonary edema. There is some retrocardiac atelectasis. No pneumothorax.

Bilateral pigtail catheters are present at the lung bas-es. Moderate right apical pneumothorax has minimal decreased since yesterday. The maximum width at the apex measures 24 cm as compared to yesterday measuring 27 cm. Opacity at the right lung base which appeared on the yesterday radiograph is more denser and is likely from an aspiration or atelectasis. Small right pleural effusion is unchanged.

As compared to the previous radiograph the patient has received a new left pleural effusion. The effusion on the left is constant. The extent of the pre-existing pleural effusion is constant. Unchanged appearance of the cardiac silhouette. Unchanged position of the left picc line.

As compared to the previous radiograph the right pleural effusion has minimally decreased. The extent of the pleural effusion on the right is constant. There is a right lung base opacity. The presence of a minimal right apical pneumothorax cannot be excluded. Unchanged size of the cardiac silhouette. Unchanged position of the left picc line.

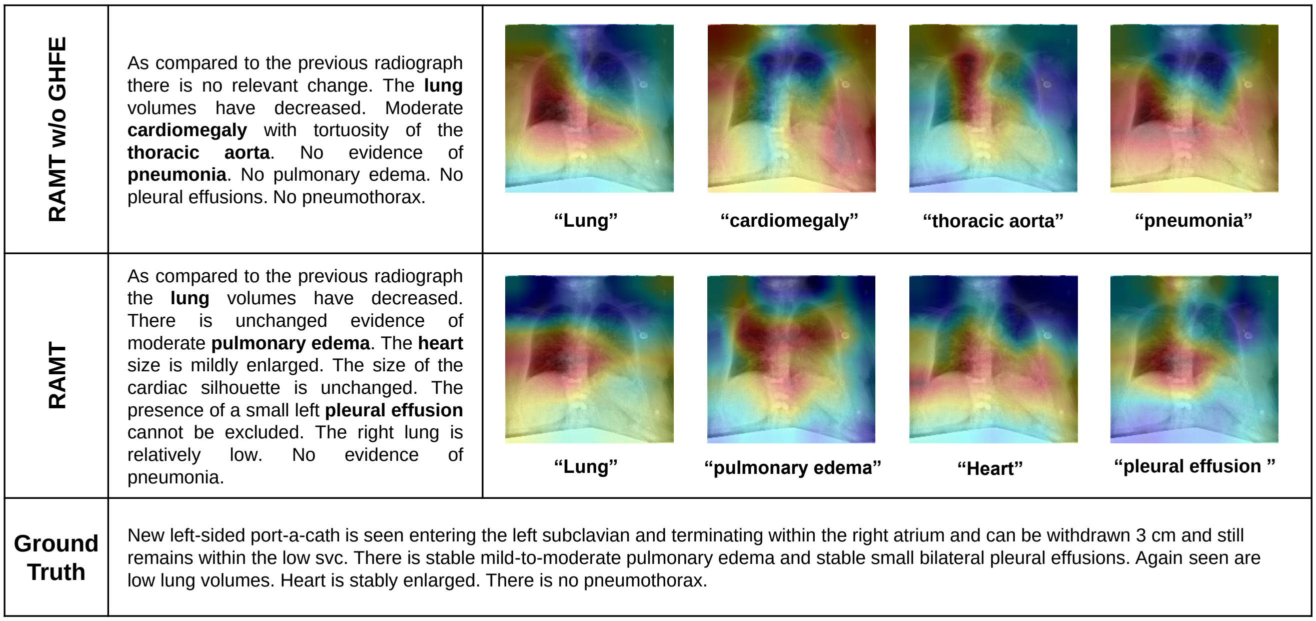
Fig. 4. Visualization and reports comparison between our RAMT framework and RAMT w/o GHFE. The blue and purple colored texts denote the generated correct sentences and wrong sentences respectively, underlined blue texts denote the correct sentences we generated but RAMT w/o GHFE did not. Missing abnormality is highlighted in Red. We additionally visualized abnormality graphs of RAMT on the right side of the chart. In the abnormality graph space, the edge weight between nodes represents their correlation strength (select the edge with the attention score greater than or equal to 0.05 for display), and the values inside parentheses correspond to the prediction probability of nodes obtained by connecting graph embedding with the extra classifier.

ROUGE-L, this indicates our consistency paradigm assists to diversify the generated reports with high recall scores. Though the BASE+GHFE with only a quarter of labeled data removes the memory-based modules in R2Gen, it still has similar performance on CE metrics to R2Gen thanks to the GHFE module that provides accurate location of the abnormality and discovers the intrinsic correlations between the local pathological changes, which will be further proved in Qualitative Analysis.

F. Ablation Study

The ablation study on MIMIC-CXR aims to investigate the contribution of each component in our proposed Graph-guided Hybrid Feature Encoding (GHFE) module. GHFE can be used in both semi-supervised and fully-supervised scenarios, the former one refers to RAMT framework and the latter one is trained without the constraints on feature correlation. In each of the scenarios, we test the effect of different components including V (only visual features are used), V+GE (visual features plus graph embedding), V+SE (visual features plus semantic embedding), GHFE (the hybrid features), and the results are shown in Table III. Obvious improvement can be observed. For example, GE and SE substantially boost the performance of basic semi-supervised framework from 0.139 to 0.144 and 0.148 respectively in METEOR score, and similar improvement about F1 score and other metrics can also be observed, which demonstrates the effectiveness of external knowledge and semantic embedding. GHFE combines the advantages of GE and SE to further improve performance. For fully supervised training, it is clear that GE and SE also suc-

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3273390



RAMT w/o GHFE	As compared to the previous radiograph there is no relevant change. The **lung** volumes have decreased. Moderate **cardiomegaly** with tortuosity of the **thoracic aorta**. No evidence of **pneumonia**. No pulmonary edema. No pleural effusions. No pneumothorax.	"Lung" "cardiomegaly" "thoracic aorta" "pneumonia"
RAMT	As compared to the previous radiograph the **lung** volumes have decreased. There is unchanged evidence of moderate **pulmonary edema**. The **heart** size is mildly enlarged. The size of the cardiac silhouette is unchanged. The presence of a small left **pleural effusion** cannot be excluded. The right lung is relatively low. No evidence of pneumonia.	"Lung" "pulmonary edema" "Heart" "pleural effusion "
Ground Truth	New left-sided port-a-cath is seen entering the left subclavian and terminating within the right atrium and can be withdrawn 3 cm and still remains within the low svc. There is stable mild-to-moderate pulmonary edema and stable small bilateral pleural effusions. Again seen are low lung volumes. Heart is stably enlarged. There is no pneumothorax.	

Fig. 5. Visualizations of image-text mappings between particular regions of a chest X-Ray image and bold words/phrases from its reports generated by RAMT w/o GHFE and RAMT, respectively. The intensity of the mappings is illustrated on the images with different colors.

cessfully boost the performance of the basic backbone network in Fig. 3(a) with only visual features, verifying the capability of our approach to enhance the basic backbone network. More detailed experiments on variants of our proposed method are provided in Table AIII in the Appendix.

G. Qualitative Analysis

To intuitively demonstrate the performance of the GHFE module in our RAMT framework, we compare the reports generated by RAMT with those generated by RAMT without (w/o for short) GHFE (the semi-supervised learning from only visual features using report consistency loss and supervised loss) and ground truth reports. The visualization and reports with abnormality graphs are shown in Fig. 4. The reports generated by our RAMT are more accurate and contain more semantic details than RAMT w/o GHFE, and tend to concern more diversified visual regions. Specifically, in the first sample, RAMT correctly detected abnormalities with correct positions such as ”left pleural effusions” and ”atelectasis” from truncated images with limited costal partial angle, which are ignored by RAMT w/o GHFE. From the doctor’s viewpoint, pleural edema is often accompanied by atelectasis, and both of them can be caused by same kind of disease e.g., pneumonia or central primary lung cancer. The first abnormality graph shows that the edge weight between ”Edema” and ”Atelectasis” is 0.129, which is much higher than the threshold of 0.05, further indicating that there is a correlation between the two. Accordingly, we can observe the high response degree of attention to the junction of left lung and heart from the attention map. We think GHFE makes full use of medical prior knowledge to locate the potential positions where abnormalities may occur and capture the relationship between lung pathological changes. However, some abnormalities, e.g., cardiomegaly, are still missing (highlighted in red). The reason may be that the model focuses more on the lung abnormalities, but less on the heart mediastinum that is relatively independent of lung part. This can be corroborated in the first abnormality graph, in which there is no edge with strong correlation connecting lung diseases and cardiomegaly. RAMT w/o GHFE misjudged the cardiac status as ”constant appearance of the cardiac silhouette” which is consistent with its attention map where the focus area is the upper lung. Without knowledge support, RAMT w/o GHFE tends to take heuristic measures, e.g. using phrases with high word frequency, leading to incorrect report generation. In the second sample, our RAMT successfully predicted pneumothorax and pleural effusion by fully capturing the intrinsic correlations between pleural-related pathological changes, which may be caused by trauma or tuberculosis. Furthermore, it can also be observed in the second abnormality graph that the edge weight between ”Effusion” and ”Pneumothorax” is 0.103 much higher than the threshold. This further validates the ability of RAMT to capture the intrinsic relations of abnormalities. On the contrary, RAMT w/o GHFE missed some abnormalities because of the inability to grasp the correlations between the local pathological changes.

H. More Case Studies

In order to further verify the effectiveness of our approach, we perform a case study on RAMT w/o GHFE and RAMT with an example input chest X-Ray image chosen from the test set of MIMIC-CXR. We visualize image-text mappings between particular regions of the chest X-Ray image and keywords/phrases from its reports. As shown in Fig. 5, it

TABLE IV
EFFECT OF β FOR RAMT ON MIMIC-CXR DATASET.

β	0.05	0.2	1	5	20
B-1	0.346	**0.358**	0.355	0.354	0.337
B-2	0.210	**0.221**	0.216	0.213	0.205
M	0.142	**0.153**	0.148	0.145	0.136
R-L	0.277	**0.289**	0.284	0.282	0.272

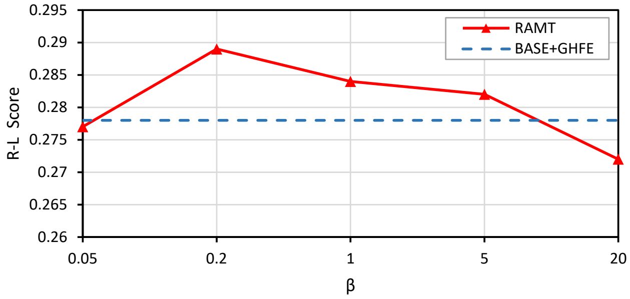


Fig. 6. Effect of varying β on R-L score via line chart built from Table IV.

can be observed that RAMT w/o GHFE tends to pay extra attention to redundant abdomen and lateral areas, interfering with the judgment of disease, while RAMT focuses more accurately than RAMT w/o GHFE on locating the key regions through hybrid features. For instance, when RAMT w/o GHFE locates the region of the word ”cardiomegaly”, in addition to the location of the heart, it also locates more redundant surrounding areas, while RAMT locates the heart region more accurately biased towards the center. This demonstrates that the GHFE module indeed enhances the robustness of our RAMT to learn image-text mappings between particular image regions.

I. Sensitivity Analysis

Furthermore, we also study the sensitivity of different hyperparameter settings for β and report the performance of our RAMT framework under 25% labeled pairs. As shown in Table IV and Fig. 6, it can be observed that the performance is not very sensitive to the value of β in the range from 0.2 to 5. However, when the value of β comes to 20, the R-L score decreases significantly. The reason being that although λ is negligible at the beginning of model training, the high value of β enlarges ratio of the intermediate consistency loss, which may introduced excessive interference before the model parameters become stabilized, resulting in the convergence to the local optimal value. As we can see, the best performance is achieved when β equals 0.2. Thus, the value of β is set as 0.2 in our experiments.

V. Conclusion

In this paper, we propose a semi-supervised framework named as Relation-Aware Mean Teacher (RAMT) for medical report generation with insufficient labeled reports. RAMT achieves the semi-supervised learning using a mean teacher paradigm. Owing to the proposed Graph-based Hybrid Features Encoding (GHFE) module, RAMT can exploit graph-form external disease knowledge to discover the graph embedding and semantic embedding and combine them with visual features for report generation. Extensive experiments on two widely-used public benchmark datasets demonstrate the superiority of our approach using only a few labeled pairs. The proposed framework is open-ended, which means the backbone network of our framework can be replaced with any other more powerful structure e.g. PVT [43], GPT-3 [44] than CNN-Transformer to reach better performance.

References

[1] L. Delrue, R. Gosselin, B. Ilsen, A. V. Landeghem, J. d. Mey, and P. Duyck, “Difficulties in the interpretation of chest radiography,” in *Comparative interpretation of CT and standard radiography of the chest*. Springer, 2011, pp. 27–49.

[2] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, “Multitask learning for cross-domain image captioning,” *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2018.

[3] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, “Exploring pairwise relationships adaptively from linguistic context in image captioning,” *IEEE Transactions on Multimedia*, 2021.

[4] L. Yu, J. Zhang, and Q. Wu, “Dual attention on pyramid feature maps for image captioning,” *IEEE Transactions on Multimedia*, vol. 24, pp. 1775–1786, 2021.

[5] Q. Huang, Y. Liang, J. Wei, Y. Cai, H. Liang, H.-f. Leung, and Q. Li, “Image difference captioning with instance-level fine-grained feature representation,” *IEEE Transactions on Multimedia*, vol. 24, pp. 2004–2017, 2021.

[6] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang, and T. Mei, “Unpaired image captioning with semantic-constrained self-learning,” *IEEE Transactions on Multimedia*, vol. 24, pp. 904–916, 2021.

[7] Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Hybrid retrieval-generation reinforced agent for medical image report generation,” *Advances in neural information processing systems*, vol. 31, 2018.

[8] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Knowledge-driven encode, retrieve, paraphrase for medical image report generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6666–6673.

[9] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1439–1449.

[10] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, “When radiology report generation meets knowledge graph,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 910–12 917.

[11] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, “Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs,” *arXiv preprint arXiv:1901.07042*, 2019.

[12] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, “Preparing a collection of radiology examinations for distribution and retrieval,” *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[13] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[15] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, “Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.

[16] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 457–466.

[17] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 721–729.

[18] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2577–2586.

[19] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest x-ray reports," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6570–6580.

[20] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 448–458.

[21] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5904–5914.

[22] F. Liu, X. Wu, S. Ge, W. Fan, and Y. 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*, 2021, pp. 13 753–13 762.

[23] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.

[24] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3942–3951.

[25] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Machine Intelligence*, vol. 4, no. 1, pp. 32–40, 2022.

[26] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme *et al.*, "Learning to exploit temporal structure for biomedical vision-language processing," *arXiv preprint arXiv:2301.04558*, 2023.

[27] H.-Y. Zhou, C. Lian, L. Wang, and Y. Yu, "Advancing radiograph representation learning with masked record modeling," *arXiv preprint arXiv:2301.13155*, 2023.

[28] D. Wang, Y. Zhang, K. Zhang, and L. Wang, "Focalmix: Semi-supervised learning for 3d medical image detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3951–3960.

[29] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8801–8809.

[30] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 490–498.

[31] A. I. Aviles-Rivero, N. Papadakis, R. Li, P. Sellars, Q. Fan, R. T. Tan, and C.-B. Schönlieb, "Graphx net-chest x-ray classification under extreme minimal supervision," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI*, 2019, pp. 504–512.

[32] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[33] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[34] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[35] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3429–3440, 2020.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[37] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[39] S. Banerjee and A. 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*, 2005, pp. 65–72.

[40] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[41] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. 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*, 2018, pp. 6077–6086.

[42] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[43] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

Ke Zhang received the B.Eng and M.Sc. degrees from Hangzhou Dianzi University, Hangzhou, China. He is currently working toward the Ph.D. degree with the Key Laboratory of Complex Systems Modeling and Simulation in the School of Computer Science, Hangzhou Dianzi University, Hangzhou, China. His research interests include deep learning, cross-modal learning, and medical image analysis.

Hanliang Jiang received the bachelor's degree in fundamental medicine and the Ph.D. degree in internal medicine from Zhejiang University, Zhejiang, China. He is studying at lung cancer mechanism and comprehensive therapy. He is currently a Physician in the Pulmonary and Critical Care Medicine department, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China.

Jian Zhang received the Ph.D. degree from Zhejiang University, Zhejiang, China. He is currently a Professor with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China. From 2009 to 2011, he was with Department of Mathematics of Zhejiang University as a Post-doctoral Research Fellow. In 2016, he had been doing research on machine learning at Simon Fraser University (SFU) as a Visiting Scholar. His research interests include but are not limited to machine learning and multimodal learning. He served as a reviewer of several prestigious journals including IEEE Transactions on Image Processing and IEEE Transactions on Circuits and Systems for Video Technology.

Qingming Huang (Fellow, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Chair Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 400 academic papers in prestigious international journals, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, and IEEE Transactions on Circuits and Systems for Video Technology, and top-level conferences, such as NIPS, ICCV, CVPR, ACM Multimedia, IJCAI, AAAI, and VLDB. His research interests include multimedia computing, image processing, computer vision, and pattern recognition. He is an Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology and Acta Automatica Sinica, and a Reviewer of various international journals, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and IEEE Transactions on Multimedia. He has served as the General Chair, Program Chair, Track Chair, and TPC Member for various conferences, including the ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, and PSIVT.

Jianping Fan is a professor at UNC-Charlotte. He received his MS degree in theory physics from Northwestern University, Xian, China in 1994 and his PhD degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997. He was a Postdoc Researcher at Fudan University, Shanghai, China, during 1997-1998. From 1998 to 1999, he was a Researcher with Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From 1999 to 2001, he was a Postdoc Researcher in the Department of Computer Science, Purdue University, West Lafayette, IN. His research interests include image/video privacy protection, automatic image/video understanding, and large-scale deep learning.

Jun Yu (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China. He was an Associate Professor with the School of Information Science and Technology, Xiamen University, Xiamen, China. From 2009 to 2011, he was with Nanyang Technological University, Singapore. From 2012 to 2013, he was a Visiting Researcher with Microsoft Research Asia (MSRA). He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. He has authored or coauthored more than 100 scientific articles. Over the past years, his research interests have included multimedia analysis, machine learning, and image processing. In 2017, he received the IEEE SPS Best Paper Award. He has (co-)chaired several special sessions, invited sessions, and workshops. He has served as a program committee member for top conferences including CVPR, ACM MM, AAAI, IJCAI, and has served as associate editors for prestigious journals including IEEE Trans. CSVT and Pattern Recognition.

Weidong Han received his bachelor's degree and Ph.D. from the College of Medicine at Zhejiang University in Zhejiang, China. He currently serves as a physician in the Department of Medical Oncology at Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University in Hangzhou, China. Additionally, he is a professor in the College of Mathematical Medicine at Zhejiang Normal University in Jinhua, China. With over 100 published articles and more than 3000 citations, his research interests center around the application of imaging omics in tumor diagnosis and treatment. He has served as a reviewer for several prestigious journals, including Advanced Science, Biomaterials, Autophagy, Signal Transduction and Targeted Therapy, and Theranostics etc.

APPENDIX

This appendix gives the detailed interpretation of notations and the definitions in this paper listed in Table AI, the characteristics of the two datasets in Table AII, and more detailed experiments on variants of our proposed method are provided in Table AIII.

TABLE AI
MAIN NOTATIONS AND THE DEFINITIONS.

Notations	**Definitions**
D_L, D_U	The labeled dataset and the unlabeled dataset.
N, M	The number of labeled samples (i.e., D_L) and the number of unlabeled samples (i.e., D_U).
θ_s^k, θ_t^k	The parameters of student and teacher network in k-th training iteration.
$\mathbf{p}_i$, $\mathbf{q}_i$	The report predicted by student and teacher corresponding to image x_i.
$\mathbf{x}_i^{\tau_s}$, $\mathbf{x}_i^{\tau_t}$	Image x_i perturbed by noise τ_s and τ_t respectively.
$\mathbf{h}_s$, $\mathbf{h}_t$	Vectorized intermediate representation from the student and teacher intermediate layers.
$\tilde{\mathrm{F}}_s(\cdot)$, $\tilde{\mathrm{F}}_t(\cdot)$	Student and teacher subnetworks including GHFE module to generate the hybrid features.
$\mathcal{L}_S$, $\mathcal{L}_U$, $\mathcal{L}_R$, $\mathcal{L}_I$	The supervised loss, overall unsupervised loss, report consistency loss and intermediate consistency loss respectively.
λ, β	Trade-off weights to balance the ratio of different losses.
$\mathbf{V}_f$, $\mathbf{V}_g$, $\mathbf{V}_s$	The visual features, the Graph Embedding features and the Semantic Embedding features.
$\mathbf{G}^0$, $\tilde{\mathbf{G}}$, $\mathbf{G}^l$	Initial graph node features, intermediate state of graph node features, graph node features from l-th graph convolutional layer.
$\mathbf{S}$, $\tilde{\mathbf{S}}$	Semantic features retrieved from learned word dictionary, and simplified semantic features with zero lines removed.
$\mathbf{V}_{Hybrid}$	Hybrid feature representation consisting of $\mathbf{V}_f$, $\mathbf{V}_g$ and $\mathbf{V}_s$.
$\mathbf{H}_{f/g/s}$	The gram matrix reflecting the internal relationship from any one of the hybrid features.
$\mathbf{H}$	Concatenation of the three gram matrices (i.e., $\mathbf{H}_f$, $\mathbf{H}_g$ and $\mathbf{H}_s$) with normalization.

TABLE AII
THE STATISTICAL INFORMATION OF THE TWO DATASETS.

Dataset	**IU X-Ray**			**MIMIC-CXR**		
	Train	Validate	Test	Train	Validate	Test
Image	5.2k	0.7k	1.5k	369.0k	3.0k	5.2k
Report	2.8k	0.4k	0.8k	222.8k	1.8k	3.3k
Patient	2.8k	0.4k	0.8k	64.6k	0.5k	0.3k
Avg Len	37.6	36.8	33.6	53.0	53.1	66.4

TABLE AIII
COMPARATIVE EXPERIMENTS ON IU X-RAY AND MIMIC-CXR DATASET, WHERE THE RESULTS ABOUT NLG METRICS OF OUR PROPOSED METHOD AND ITS VARIANTS WITH DIFFERENT RATIOS OF LABELED TRAINING SAMPLES ARE SHOWN. THE DIFFERENCE BETWEEN THE UPPER PART AND THE LOWER PART IS WHETHER THE GHFE MODULE WE PROPOSED IS INCLUDED.

Method	Ratio		IU X-Ray						MIMIC-CXR					
	L	U	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
BASE	25	0	0.406	0.257	0.183	0.136	0.169	0.354	0.313	0.192	0.127	0.091	0.128	0.270
RAMT w/o GHFE	25	75	0.458	0.279	0.196	0.142	0.181	**0.362**	0.335	0.207	0.140	0.097	0.139	0.276
Upperbound	100	0	**0.466**	**0.295**	**0.205**	**0.148**	**0.183**	0.359	**0.348**	**0.216**	**0.144**	**0.101**	**0.144**	**0.278**
BASE+GHFE	25	0	0.433	0.278	0.197	0.148	0.176	0.354	0.323	0.202	0.136	0.097	0.134	0.278
RAMT	25	75	0.480	0.302	0.214	0.159	**0.196**	0.368	0.358	0.221	0.148	0.106	**0.153**	**0.289**
RAMT-Upperbound	100	0	**0.482**	**0.310**	**0.221**	**0.165**	0.195	**0.377**	**0.362**	**0.229**	**0.157**	**0.113**	**0.153**	0.284