Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation

Yan Yang, *Student Member, IEEE*, Jun Yu, *Member, IEEE*, Jian Zhang, Weidong Han, Hanliang Jiang, and Qingming Huang, *Fellow, IEEE*

***Abstract*—Medical image report generation (MeIRG) aims at generating associated diagnosis descriptions with natural language sentences from medical images, which is essential in the computer-aided diagnosis system. Nevertheless, this task remains challenging in that medical images and linguistic expressions should be understood jointly which however show great discrepancies in the modality. To fill this visual-to-semantic gap, we propose a novel framework that follows the encoder-decoder pipeline. Our framework is characterized by encoding both deep visual and semantic embeddings through a triple-branch network (TriNet) during the encoding phase. The visual attention branch captures attended visual embeddings from medical images with the soft-attention mechanism. The medical report (MeRP) embedding branch predicts semantic report embeddings. The embedding branch of medical subject headings (MeSH) obtains semantic embeddings of related medical tags as complementary information. Then, outputs of these branches are fused and fed into a decoder for the report generation. Experimental results on two benchmark datasets have demonstrated the excellent performance of our method. Related codes are available at https://github.com/yangyan22/Medical-Report-Generation-TriNet.**

***Index Terms*—Medical image report generation, visual-semantic joint embedding, encoder-decoder framework, deep neural network.**

Manuscript received 15 January 2021; revised 16 July 2021; accepted 19 October 2021. Date of publication 8 November 2021; date of current version 13 January 2023. This work was supported in part by the National Natural Science Foundation for Distinguished Young Scholars of China under Grant 62125201 and in part by the National Natural Science Foundation of China under Grant 62176230. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Jianguo Zhang. (Corresponding authors: Jun Yu; Weidong Han.)

Yan Yang and Jun 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: yangyanyy@hdu.edu.cn; yujun@hdu.edu.cn).

Jian Zhang is with the Zhejiang International Studies University, Hangzhou 310023, China (e-mail: jeyzhang@outlook.com).

Weidong Han is with the Department of Medical Oncology, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou 310016, China (e-mail: hanwd@zju.edu.cn).

Hanliang Jiang is with the 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).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3122542.

Digital Object Identifier 10.1109/TMM.2021.3122542

I. Introduction

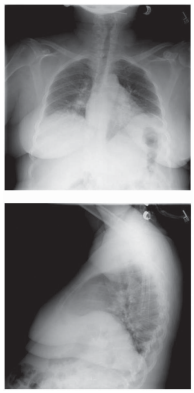
MEDICAL imaging has been widely employed in clinical practice to assist disease diagnosis. However, medical image reading and report writing can be time-consuming and knowledge-demanding. This motivates researchers to study medical image report generation (MeIRG), i.e., automatically generating diagnosis descriptions with natural language sentences based on the given medical images. Fig. 1 shows a pair of frontal and lateral chest X-ray images associated with the medical report (MeRP) and the medical subject headings (MeSH) that record important medical tags. The MeRP is composed of the *impression* and *findings*, where the *findings* describe detailed information for both the normal and abnormal observations of different body parts while the *impression* summaries *findings* with a diagnostic result.

The earliest works [1], [2] in the field of MeIRG tried to attach relevant "texts" to medical images. In these methods, target "texts" are either fully-structured or semi-structured (e.g. tags, templates), rather than the unstructured linguistic expressions like human observers. Therefore, these methods are not able to generate diverse and flexible diagnostic reports with rich contextual information. Recently, encoder-decoder architectures based on deep learning have shown advantages in image captioning [3]–[5]. These methods employ an encoder to encode images with feature vectors and generate corresponding captions with a decoder. Inspired by the great success of image captioning, most existing MeIRG methods follow the encoder-decoder pipeline [6]–[8] to generate unstructured free-text reports. But MeIRG can be different from image captioning that aims at generating only one summative sentence to describe the main content. In the task of MeIRG, a lengthy paragraph consisting of fine-grained sentences should be generated to describe detailed observations of different physical tissues. Since building the relationship between visual contents of medical images and linguistic expressions is difficult, automatically generating detailed and accurate reports from medical images remains challenging.

To fill the visual-to-semantic gap, we attempt to obtain both the visual information and the complementary semantic information for MeIRG. The encoder is composed of three branches, including MeRP embedding branch, MeSH embedding branch, and visual attention branch. The MeRP embedding branch and the MeSH embedding branch aim at predicting semantic embeddings by exploiting semantic term frequency-inverse document frequency (TF-IDF) [9] features from the MeRP and MeSH corpus respectively. The visual attention branch explicitly learns the

1520-9210 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.



Impression:
Low lung volumes with right basilar atelectasis. Otherwise, no acute cardiopulmonary disease.

Findings:
There extremely low lung volumes. There is right basilar opacity. There is no pneumothorax. There is no large pleural effusion. Cardiac silhouette and mediastinal contours are within normal limits.

MeSH:
Atelectasis, Opacity

Fig. 1. An example of the chest X-ray image pairs associated with the corresponding MeRP (including the *impression* and *findings* sections) and MeSH. The upper image is of the frontal view and the bottom image is of the lateral view.

attended visual contents of medical images with the soft attention mechanism [5]. Then, we fuse the outputs of these branches by a topic encoding module with the long short-term memory (LSTM) [10] to obtain holistically represented topics, which are then fed to a decoding module for generating sentences in the report.

Overall, the main contributions of our paper can be summarized as:

1) We propose a triple-branch network (TriNet) to encode both visual contents and complementary semantic information for detailed and accurate MeIRG. The semantic embeddings are achieved by predicting TF-IDF features of both the MeRP and MeSH.
2) Our framework is implemented via an end-to-end multi-task network. The encoding phase generates fused visual-semantic embeddings as well as the MeSH information, and the decoding phase generates *findings* and the *impression* sections simultaneously.
3) Comprehensive experiments in two publicly available datasets have justified the excellent performance of the proposed architecture and have demonstrated the significance of each branch.

II. Related Works

In this section, we review related works in the field of image captioning and medical image report generation.

A. Image Captioning

Image captioning aims at generating a summative description for the input image automatically. The earliest works generate image captions with template-based or retrieval-based methods [11]. Template-based approaches [12], [13] usually pre-define fixed templates with several blank slots to be filled. In these methods, objects, attributes, relationships, and actions will be classified to fill the blank slots in templates. Although template-based approaches can generate grammatically correct captions, they are not able to generate variable-length captions due to the structured templates. Retrieval-based approaches [14], [15] retrieve captions for the query image from a caption pool. Retrieval-based methods could generate syntactically correct captions, but they are not able to generate image-specific captions.

In recent years, deep learning models have achieved excellent performance in image captioning [16]–[21]. Vinyals *et al.* [3] proposed an encoder-decoder deep-learning framework for image captioning, which combined advances in computer vision and machine translation. Xu *et al.* [5] introduced an attention mechanism for image caption to dynamically attend to semantic-related visual contents and significantly improved the performance. You *et al.* [22] proposed to employ semantic attention for image captioning, which selectively attended to semantic proposals. The top-down methods extract visual features first and then convert these features into words. The bottom-up methods first extract visual concepts (e.g., objects, relationships, and attributes) from the image and then combine them for captioning. The aforementioned methods typically predict the next word according to previous words and visual features [3]–[5]. If the previous words were predicted wrong, the error would accumulate. To generate natural and fluent captions, reinforcement learning (RL) [18], [23] based, and generative adversarial networks (GAN) based [24], [25] methods are further proposed. Most recently, the scene-graph-based models for image captioning are proposed [26], [27].

B. Medical Image Report Generation

Deep learning-based approaches are currently dominating the solutions to the MeIRG task. Shin *et al.* [28] adopted convolutional neural networks (CNN) and recurrent neural networks (RNN) to generate object labels and describe their context like human observers. Jing *et al.* [6] employed a hierarchical LSTM framework to generate reports and predicted related semantic tags to assist report generation with a co-attention mechanism. Wang *et al.* [8] proposed a text-image embedding network, which extracts visual features by CNN and generates reports word-by-word with an LSTM. The attention-encoded text embedding and saliency weighted global pooling are used for disease classification. Xue *et al.* [29] employed a recurrent model for the report generation, in which a start sentence was firstly generated and then the succeeding sentence was generated by taking the preceding sentence and the visual information as input. Li *et al.* [30] proposed a hybrid retrieval-generation reinforced agent for MeIRG, which reconciled retrieval-based approaches with the reinforcement learning methods to achieve structured, robust, and diverse report generation. Li *et al.* [31] proposed knowledge-driven encoding, retrieving, paraphrasing for MeIRG, where they applied both knowledge-based and retrieval-based approaches with the learning-based methods. Jing *et al.* [32] proposed a two-stage strategy to model the relationship between *findings* and *impression*, and they designed a cooperative multi-agent system to deal with the imbalanced data distribution between the abnormality and normality. Harzig *et al.* [33] proposed to address data bias problems for report generation. They separated the abnormal and normal sentence generation using a dual word LSTM with a hierarchical LSTM model. Yuan *et al.* [34] incorporated multi-view visual contents with sentence-level attention and enforced the consistency between

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

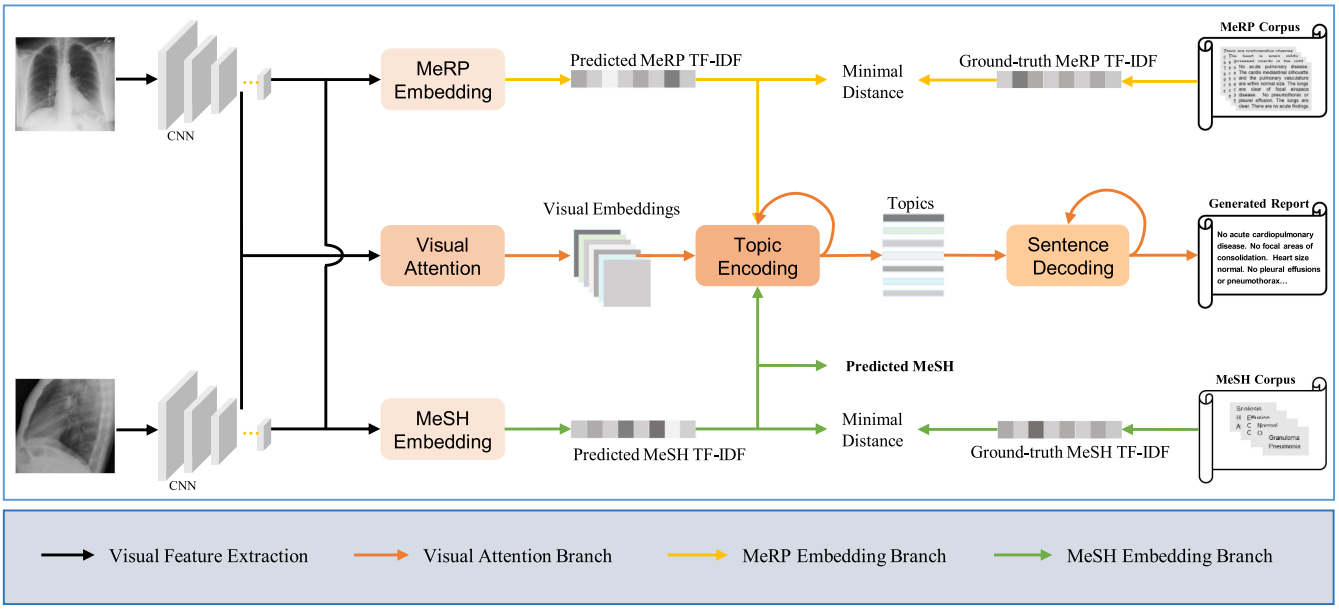


Fig. 2. Workflow of the proposed TriNet for medical report generation. First, visual features from multi-view medical images are extracted with different CNNs which share the same architecture (ResNet-50 [40]). Second, visual embeddings for topic encoding are obtained by the visual attention mechanism. Then, visual-to-semantic embeddings are captured by predicting corresponding TF-IDF features of both the MeRP and MeSH within the MeRP and the MeSH corpus. After that, these embeddings are jointly encoded into fused topics. Finally, all the topics are fed to a sentence decoder to generate the intact report.

different views for report generation. They also extracted medical concepts to assist report generation with word-level attention. Xie *et al.* [35] proposed an attention-based abnormal-aware network for report generation, where they separated sentence generation into abnormal and normal sentence generation through a high-level gate module. Zhang *et al.* [7] proposed to utilize a pre-constructed graph embedding on multiple disease findings to assist report generation. They also proposed a new evaluation metric to better assess the quality of generated reports with the assistance of the proposed knowledge graph.

Compared with the aforementioned works that mostly encode visual contents in a single flow for MeIRG, our proposed TriNet tried to jointly and effectively encode both visual embeddings and semantic embeddings by utilizing a triple-branch architecture. In order to fill the cross-modality semantic gap between medical images and the reports, we exploit semantic information to assist report generation by predicting semantic TF-IDF features under the guidance of both the MeRP corpus and MeSH corpus. In this paper, we also follow some successful practices of previous works: we adopt the encoder-decoder framework to encode topics and generate associated sentences to form an intact report [6]; we follow the latest fashion of fusing multi-view image features for the report generation as [7], [29], [35]; we also generate *findings* and *impression* sections simultaneously by a concatenation operation as [7], [8].

III. Methodology

In this section, we present the proposed methodology. Fig. 2 shows the framework of TriNet, which encodes multiple topics and decodes the topics to generate associated sentences in the report. Firstly, visual features from multi-view medical images are extracted with different CNNs and the visual embeddings for topic encoding are obtained with the visual attention mechanism. Then, visual-to-semantic embeddings are obtained by predicting corresponding TF-IDF features of the MeRP and MeSH within the MeRP and the MeSH corpus. After that, these embeddings are jointly encoded into fused topics. Finally, all these topics are fed to a sentence decoder module to generate corresponding sentences. The main notations and the definitions in this paper are listed in Table I.

A. Visual Attention Branch

The visual attention branch aims at obtaining visual embeddings from multi-view images in a medical study for MeIRG. In medical imaging diagnosis, each sentence in the report usually has unique visual evidence. In the meanwhile, sentences in a single report usually have strong contextual dependencies, which should also be reflected in the encoded features. Therefore, we adopt an attention mechanism to attend to different visual contents of the frontal and lateral views for topic encoding at each time step of the encoding LSTM.

Specifically, we first extract frontal visual features $\mathbf{v}_{f,n}$ and lateral visual features $\mathbf{v}_{l,n}$ from an intermediate layer of the CNNs where $n \in (1, N)$ as shown in Fig. 2 (i.e., visual feature extraction). Then, we feed them and the previous hidden state $\mathbf{h}_{encoder}^{(s-1)}$ of the encoding LSTM into the visual attention network to obtain visual embedding $\mathbf{v}^{(s)}$ at time step s of the encoding LSTM. The details of the visual attention network are shown in Fig. 3, which can be formulated with the following equations. First, attention weights over frontal and lateral visual features will be obtained by:

$$\alpha_{f,n}^{(s)} \propto \exp(\mathbf{W}_{f_{att}} \tanh(\mathbf{W}_{f,v}\mathbf{v}_{f,n} + \mathbf{W}_{f,h}\mathbf{h}_{encoder}^{(s-1)}))$$
$$\alpha_{l,n}^{(s)} \propto \exp(\mathbf{W}_{l_{att}} \tanh(\mathbf{W}_{l,v}\mathbf{v}_{l,n} + \mathbf{W}_{l,h}\mathbf{h}_{encoder}^{(s-1)})) \quad (1)$$

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

TABLE I
MAIN NOTATIONS AND THE DEFINITIONS

Notation	Definition
$\mathbf{v}_{f,n}$, $\mathbf{v}_{l,n}$	Frontal visual features, and lateral visual features extracted from an intermediate layer of CNN.
$\mathbf{v}_f^{(s)}$, $\mathbf{v}_l^{(s)}$	The attended frontal visual feature, and the attended lateral visual feature at time step s of the encoding LSTM.
$\mathbf{v}^{(s)}$	Visual embedding for topic encoding at time step s of the encoding LSTM in the topic encoding module.
$\mathbf{s}^r$, $\mathbf{s}^m$	Ground-truth TF-IDF vector of the MeSH, and ground-truth TF-IDF vector of the MeRP.
$\hat{\mathbf{s}}^r$, $\hat{\mathbf{s}}^m$	Predicted TF-IDF vector of the MeSH, and predicted TF-IDF vector of the MeRP.
$\mathbf{s}$, $\hat{\mathbf{s}}$	Ground-truth semantic embedding vector, and the predicted semantic embedding vector.
$\mathbf{v}_o$	Concatenated average visual feature of the frontal and lateral views from the output layer of the CNN.
$\mathbf{h}_{encoder}$, $\mathbf{c}_{encoder}$	Hidden state, and memory state of the encoding LSTM in the topic encoding module.
$\mathbf{h}_{decoder}$, $\mathbf{c}_{decoder}$	Hidden state, and memory state of the decoding LSTM in the sentence decoding module.
$\|R\|$, $\|M\|$	The number of reports in report corpus, and the number of MeSH in MeSH corpus.
$\mathbf{t}$, $\mathbf{r}$	The generated topic vector, and the ground-truth report.
$\hat{\mathbf{p}}_{stop}$, $\mathbf{p}_{stop}$	The predicted distribution, and the ground-truth distribution to determine when to stop generating topics.
$\hat{\mathbf{p}}_{word}$, $\mathbf{p}_{word}$	The predicted distribution, and the ground-truth distribution to generate words in the sentences.
$\mathbf{f}_c$, $\mathbf{f}_h$	Multi-layer perception functions for the initialization of the memory state and the hidden state of the LSTM.
ℓ_{word}, ℓ_{stop}, ℓ_{sem}	The word generation loss, the stop control loss, and the semantic embedding loss.
λ_{word}, λ_{stop} , λ_{sem}	Balancing weights of the word generation loss, the stop control loss, and the semantic embedding loss.
$\mathbf{W}$	The parameter matrix.

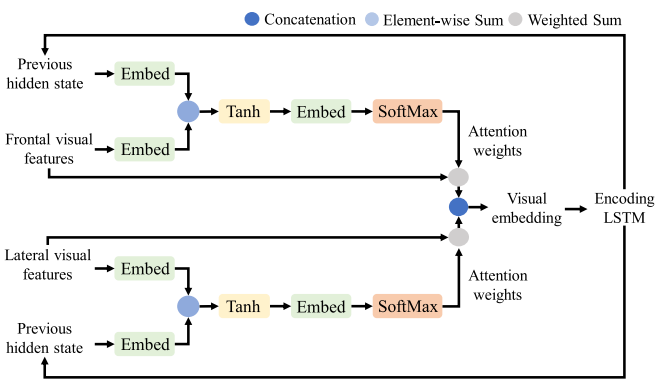


Fig. 3. Details of the visual attention network. First, visual features of frontal or lateral images and the previous hidden state of the encoding LSTM are embedded and added. Then, the added feature will be activated by a Tanh and further embedded. After that, a SoftMax activation function is employed to obtain attention weights which will be used to attend different visual features by a weighted sum operation. Finally, the attended frontal and lateral features will be concatenated as the visual embedding for topic encoding.

where $\mathbf{W}_{f,v}$, $\mathbf{W}_{f,h}$, and $\mathbf{W}_{f_{att}}$ are parameter matrices for frontal visual attention. $\mathbf{W}_{l,v}$, $\mathbf{W}_{l,h}$, and $\mathbf{W}_{l_{att}}$ are parameter matrices for lateral visual attention.

Then, we compute the attended frontal and lateral visual contents with the obtained attention weights as:

$$\mathbf{v}_f^{(s)} = \sum_{n=1}^{N} \alpha_{f,n}^{(s)} \mathbf{v}_{f,n}, \qquad \mathbf{v}_l^{(s)} = \sum_{n=1}^{N} \alpha_{l,n}^{(s)} \mathbf{v}_{l,n} \tag{2}$$

Finally, the attended visual contents $\mathbf{v}_f^{(s)}$ and $\mathbf{v}_l^{(s)}$ are concatenated as visual embedding $\mathbf{v}^{(s)} = [\mathbf{v}_f^{(s)}; \mathbf{v}_l^{(s)}]$, in which $[;]$ is the operation of concatenation.

B. MeRP Embedding Branch

Cross-modality semantic gaps between image contents and linguistic expressions have been the main challenge for MeIRG. Therefore, we attempt to exploit semantic information from the MeRP corpus to assist report generation. So far, statistical TF-IDF algorithm has been proved effective to extract keywords and topics of a text within a textual corpus, which is composed of the term frequency (TF) and inverse document frequency (IDF). TF extracts high-frequency words from the input text as candidate keywords while IDF imposes weights on TF to extract the most important words.

In this work, we predict semantic MeRP embedding $\hat{\mathbf{s}}^r$ from visual inputs by minimizing the distance between the output of MeRP embedding branch and the ground-truth MeRP TF-IDF. We first define the vocabulary of MeRP corpus. Then, for each report, we construct a ground-truth TF-IDF vector $\mathbf{s}^r$ whose element $\mathbf{s}_{i,j}^r$ of i_{th} word w_i in the j_{th} report r_j is computed as:

$$\mathbf{s}_{i,j}^r = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|R|}{|\{j : w_i \in r_j\}|} \tag{3}$$

where $n_{i,j}$ denotes the frequency of word w_i appeared in report r_j and $|R|$ is the number of reports in report corpus. $|\{j : w_i \in r_j\}|$ is the number of reports containing w_i.

After that, we predict semantic MeRP embedding $\hat{\mathbf{s}}^r$ with the following equation:

$$\hat{\mathbf{s}}^r = \mathbf{W}_r \mathbf{v}_o \tag{4}$$

where $\mathbf{v}_o$ is the concatenated average visual feature of frontal and lateral images from the output layer of the CNN, $\mathbf{W}_r$ is the parameter matrix for semantic report embedding.

C. MeSH Embedding Branch

Medical subject headings (MeSH) record the most important tags of medical images, which have been widely used for medical information indexing. Therefore, we exploit semantic information from the associated MeSH corpus as complementary semantic features. Similar to the semantic prediction of reports, MeSH embedding branch predicts semantic MeSH embedding $\hat{\mathbf{s}}^m$ from visual inputs. We first build a MeSH vocabulary containing the most important medical subject headings. Then, for each MeSH, we construct ground-truth TF-IDF vector $\mathbf{s}^m$ by calculating TF-IDF values for all tags in MeSH vocabulary. The

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

value $\mathbf{s}_{i,j}^m$ of i_{th} tag g_i in the j_{th} MeSH m_j is computed as:

$$\mathbf{s}_{i,j}^m = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|M|}{|\{j : g_i \in m_j\}|} \tag{5}$$

where $n_{i,j}$ is the frequency of g_i appeared in m_j and $|M|$ is the total number of MeSH in the corpus. $|\{j : g_i \in m_j\}|$ is the number of MeSH containing the tag g_i.

We then transform the concatenated average visual feature of frontal and lateral views from the output layer of the CNN with a fully connected layer to predict semantic MeSH embedding $\hat{\mathbf{s}}^m$ with the following equation:

$$\hat{\mathbf{s}}^m = \mathbf{W}_m \mathbf{v}_o \tag{6}$$

where $\mathbf{W}_m$ is the parameter matrix for semantic MeSH embedding. In addition, the framework predicts related MeSH information from $\hat{\mathbf{s}}^m$ by a threshold operation.

D. Topic Encoding

With visual and semantic embeddings, we encode different topics for the sentence decoding module to generate sentences by fusing all visual-semantic embeddings with an encoding LSTM. MeSH and MeRP embeddings contain global information of input images, and thus we use them to initialize the memory state and hidden state of the encoding LSTM in the topic encoding module. Since each generated sentence should have its visual evidence, we use attended visual embedding as the input at each time step of the encoding LSTM.

The initialization of encoding LSTM is achieved by multi-layer perception (MLP) functions $\mathbf{f}_{h,encoder}$ and $\mathbf{f}_{c,encoder}$ with semantic embeddings as:

$$\begin{aligned} \mathbf{h}_{encoder}^{(0)} &= \mathbf{f}_{h,encoder}\left(\hat{\mathbf{s}}\right) \\ \mathbf{c}_{encoder}^{(0)} &= \mathbf{f}_{c,encoder}\left(\hat{\mathbf{s}}\right) \end{aligned} \tag{7}$$

where $\hat{\mathbf{s}} = [\hat{\mathbf{s}}^r; \hat{\mathbf{s}}^m]$ denotes the concatenation of predicted MeRP and MeSH embeddings. Afterwards, we adopt the hidden states $\mathbf{h}_{encoder}^{(s)}$ of the encoding LSTM and the attended visual information $\mathbf{v}^{(s)}$ to generate topic vector $\mathbf{t}^{(s)}$ of the current step similar to [6], which can be formulated as:

$$\mathbf{t}^{(s)} = \tanh\left(\mathbf{W}_t \mathbf{h}_{encoder}^{(s)} + \mathbf{W}_v \mathbf{v}^{(s)}\right) \tag{8}$$

where the matrices $\mathbf{W}_t$ and $\mathbf{W}_v$ are parameters to encode the topic $\mathbf{t}^{(s)}$ at time step s.

In addition, we use a deep output layer to predict a distribution $\hat{\mathbf{p}}_{stop}^{(s)}$ to determine when the encoding LSTM will stop generating the next topic:

$$\hat{\mathbf{p}}_{stop}^{(s)} \propto \exp\left\{\tanh\left(\mathbf{W}_{stop} \mathbf{h}_{encoder}^{(s)}\right)\right\} \tag{9}$$

where $\mathbf{W}_{stop}$ is a parameter matrix. If distribution $\hat{\mathbf{p}}_{stop}^{(s)}$ is greater than 0.5, the encoding LSTM will stop topic encoding.

E. Sentence Decoding

The encoded topics will be further decoded to generate words of each sentence in the report by the sentence decoding module. With an encoded topic at each time step, the words in a sentence will be generated word by word with a decoding LSTM in the sentence decoding module. Specifically, we adopt the encoded topic vector $\mathbf{t}$ to initialize the memory state and hidden state of the decoding LSTM as:

$$\begin{aligned} \mathbf{h}_{decoder}^{(0)} &= \mathbf{f}_{h,decoder}\left(\mathbf{t}\right) \\ \mathbf{c}_{decoder}^{(0)} &= \mathbf{f}_{c,decoder}\left(\mathbf{t}\right) \end{aligned} \tag{10}$$

where $\mathbf{f}_{h,decoder}$ and $\mathbf{f}_{c,decoder}$ are MLP functions.

Then, a unique token START is firstly fed to the decoding LSTM, while subsequent inputs are the word sequence. In our work, the hidden state of decoding LSTM is used to predict probability distribution over all words in the vocabulary:

$$\hat{\mathbf{p}}_{word} \propto \exp\left(\mathbf{W}_{word} \mathbf{h}_{decoder}\right) \tag{11}$$

where $\mathbf{W}_{word}$ is the parameter matrix for word generation. And we select the word with maximum probability as the output at each time step of the decoding LSTM. When all the topics are decoded, we can obtain the final medical report by simple concatenation of all the sentences.

F. Parameter Training

In our approach, each training sample is a tuple $(\mathbf{I}, \mathbf{s}, \mathbf{r})$ composed of the image pairs $\mathbf{I}$, the ground-truth semantic vector $\mathbf{s} = [\mathbf{s}^r; \mathbf{s}^m]$, and the ground-truth report $\mathbf{r}$. The report contains S sentences and each sentence consists of T_s words. Given a training sample, the encoding LSTM will unroll for S times to generate S topics in combination with a stop control distribution $\mathbf{p}_{stop}$. For the decoding LSTM, given a sentence topic and a special START token, it will unroll for T_s times to generate word distribution $\mathbf{p}_{word}$ until the special token END appears. In our work, we train the whole model TriNet in an end-to-end manner. The loss function consists of three sub-losses, namely the semantic embedding loss ℓ_{sem}, the stop control loss ℓ_{stop} and the word generation loss ℓ_{word}. Specifically, the word generation loss and stop control loss are the cross-entropy (CE) loss functions, while the semantic embedding loss is the mean square error (MSE) loss function. Overall, the loss function for training can be defined as:

$$\begin{aligned} \ell(\mathbf{I}, \mathbf{s}, \mathbf{r}) = {} & \lambda_{sem} \ell_{sem}\left(\mathbf{s}, \hat{\mathbf{s}}\right) \\ & + \lambda_{stop} \sum_{s=1}^{S} \ell_{stop}\left(\mathbf{p}_{stop}^{(s)}, \hat{\mathbf{p}}_{stop}^{(s)}\right) \\ & + \lambda_{word} \sum_{s=1}^{S} \sum_{t=1}^{T_s} \ell_{word}\left(\mathbf{p}_{word}^{(s,t)}, \hat{\mathbf{p}}_{word}^{(s,t)}\right) \end{aligned} \tag{12}$$

IV. Experiments

In this section, we conduct comprehensive experiments on two publicly available datasets namely the IU X-Ray [36] and the MIMIC-CXR [37] to evaluate the performance.

A. Datasets

Indiana University Chest X-Ray Collection (IU X-Ray) [36] is a commonly used dataset for the task of MeIRG. It includes 7,470 images and 3,955 reports from 3,955 patients. The MeSH in this dataset is annotated by radiologists. The reports in this

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

TABLE II
STATISTICAL INFORMATION OF IU X-RAY AND MIMIC-CXR DATASETS

Dataset	Images	Reports	Patients	MeSH
IU X-Ray	7,470	3,955	3,955	Expert
MIMIC-CXR	377,110	227,835	65,379	Automatic

dataset including multiple sections are the free-text radiology reports from clinical practices.

Mimic Chest X-ray (MIMIC-CXR) [37] is the largest dataset for MeIRG, including 227,835 imaging studies from 65,379 patients. A total of 377,110 images are available in the dataset. The disease labels in this dataset are annotated automatically by the CheXpert [38] or the NegBio [39] labeller from the free-text radiology reports.

The statistical information of the two datasets is listed in Table II. In both datasets, a single report associated with one or more scans (usually a frontal image and a lateral image) is referred to as a study. We select studies with both frontal and lateral images, and we exclude studies without *impression* or *findings* sections. As a result, we obtain 2,939 studies (2,339 for training, 300 for validation and 300 for testing) for experiments in IU X-Ray. Similarly, we obtain 78,801 studies (62,801 for training, 8,000 for validation and 8,000 for testing) for experiments in MIMIC-CXR. In both datasets, the reports are de-identified to protect the privacy of patients. We pre-process reports by setting all letters to lower cases and removing all non-alpha tokens. For IU X-Ray, we obtain the top 30 medical subject headings as the MeSH vocabulary and the top 800 words that appeared frequently as the MeRP vocabulary. For MIMIC-CXR, we use 14 disease labels [38] as the MeSH vocabulary and obtain the top 2,000 words as the MeRP vocabulary.

B. Evaluation Metrics

In this paper, bilingual evaluation understudy (BLEU) [42], consensus-based image description evaluation (CIDEr) [43] and recall-oriented understudy for gisting evaluation (ROUGE) [44] metrics are employed for the quantitative performance assessment of the generated reports. BLEU is the most commonly used metric that measures the word overlaps between the generated medical reports and the ground-truth reports. BLEU-1, BLEU-2, BLEU-3 and BLEU-4 calculate overlaps of uni-grams, bi-grams, tri-grams and 4-grams respectively. A variety of ROUGE versions have been proposed [44], among which ROUGE-L is the most frequently used metric in MeIRG. The computation of ROUGE-L comprises two steps: firstly, calculate the ratio of the length of the longest common n-gram shared by the generated and the reference reports to the size of the reference description (Recall) and the generated description (Precision); secondly, combine the Recall and Precision to obtain ROUGE-L (F-measure). CIDEr is the cosine similarity of the n-gram TF-IDF representations between the generated and the ground-truth medical reports. The cosine similarities are calculated from uni-grams to 4-grams and their average is the CIDEr score. For MeSH prediction, we employ four metrics, the average accuracy, average precision, average specificity and the average recall across all MeSH tags to evaluate the prediction performance based on the binary classification of all disease tags.

C. Implementation Details

We train our model with PyTorch on one NVIDIA GeForce RTX 2080Ti GPU for experiments on IU X-Ray and four NVIDIA GeForce RTX 2080Ti GPUs for experiments on MIMIC-CXR. We employ two different ResNet-50 [40] pre-trained on ImageNet to extract visual features from frontal and lateral images respectively. The outputs of ResNet-50 are concatenated for semantic MeRP and MeSH embedding. We construct MeRP corpus and MeSH corpus by taking all original reports and medical subject headings written by professional clinicians from the training set. For visual attention, we extract visual features from the middle layer of the ResNet-50, which yields the feature map of $7 \times 7 \times 512$. For sentence generation, we set the dimension of hidden states and word embeddings to 512. The maximum sentences in a generated report and the maximum words in a sentence are set to 7 and 20 in both datasets. We set the batch size to 16 and 64 for IU X-Ray and MIMIC-CXR respectively. We train our model in an end-to-end manner with a learning rate 5e-5 and the Adam optimizer [41]. The early-stopping and drop-out strategies are adopted to prevent over-fitting. The balancing weights λ_{sem}, λ_{word} and λ_{stop} are set to 1, 1 and 0.5 respectively. Besides, we obtain MeSH tags from the predicted MeSH TF-IDF vector with a threshold of 0.5. It is worth noting that the aforementioned hyper-parameters are adopted by validating the performance on the validation sets.

D. Qualitative Results and Analysis

Fig. 4 and Fig. 5 show examples of the generated reports and MeSH information associated with the input frontal and lateral images in IU X-Ray and MIMIC-CXR respectively. The highlighted words are descriptions of the abnormality observations. As can be observed, our method can generate sentences that describe the normal findings accurately as shown in (c) and (d) of Fig. 4 and (d) of Fig. 5. For example, we can describe "No acute cardiopulmonary process. The lungs are clear without focal consolidation pleural effusion or pneumothorax. The cardiomediastinal silhouette is normal. No acute osseous abnormalities," which is similar to the ground-truth report "No acute cardiopulmonary process. The lungs are well inflated and clear. Heart size is normal and mediastinal contours are unremarkable. No pleural effusion or pneumothorax. Osseous structures are intact" as shown in (d) of Fig. 5. For abnormality description, our method can well describe the diseases accurately. But sometimes some diseases may be missed. For example, in (b) of Fig. 4, we generate "Emphysema" and "There are degenerative changes of the thoracic spine" accurately, but we miss the finding "There is hyperexpansion of the lungs with flattening of the hemidiaphragms". This phenomenon shows that, in MeIRG, the generation of abnormality findings will be much more difficult than the normal findings. It remains challenging to generate findings accurately for the abnormalities. The main reason could be the long-tailed distribution of training samples: some of the abnormal findings appear rarely in the dataset while the normal findings appear frequently, which could influence the prediction accuracy of the related diseases and degenerate the quality of the

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

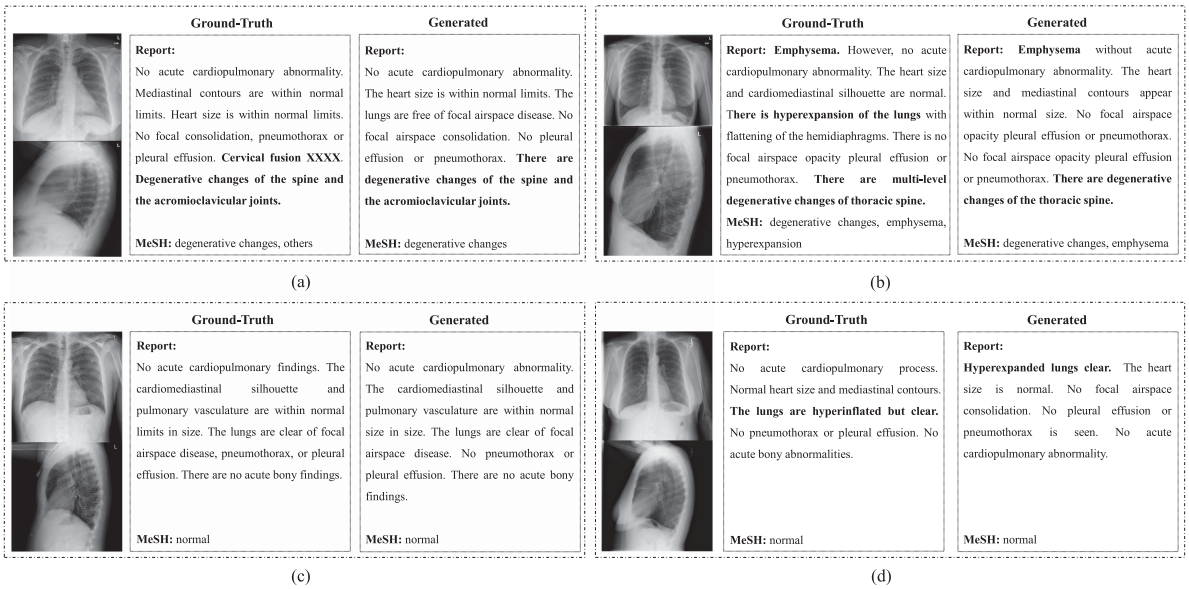


Fig. 4. Examples of generated reports given the frontal and lateral images in IU X-Ray dataset. Highlighted words are the descriptions for abnormality findings. (a) and (b) are examples with abnormalities. (c) and (d) are examples without abnormalities. The associated medical subject headings (MeSH) are presented at the bottom. "XXXX" represents the de-identified word.

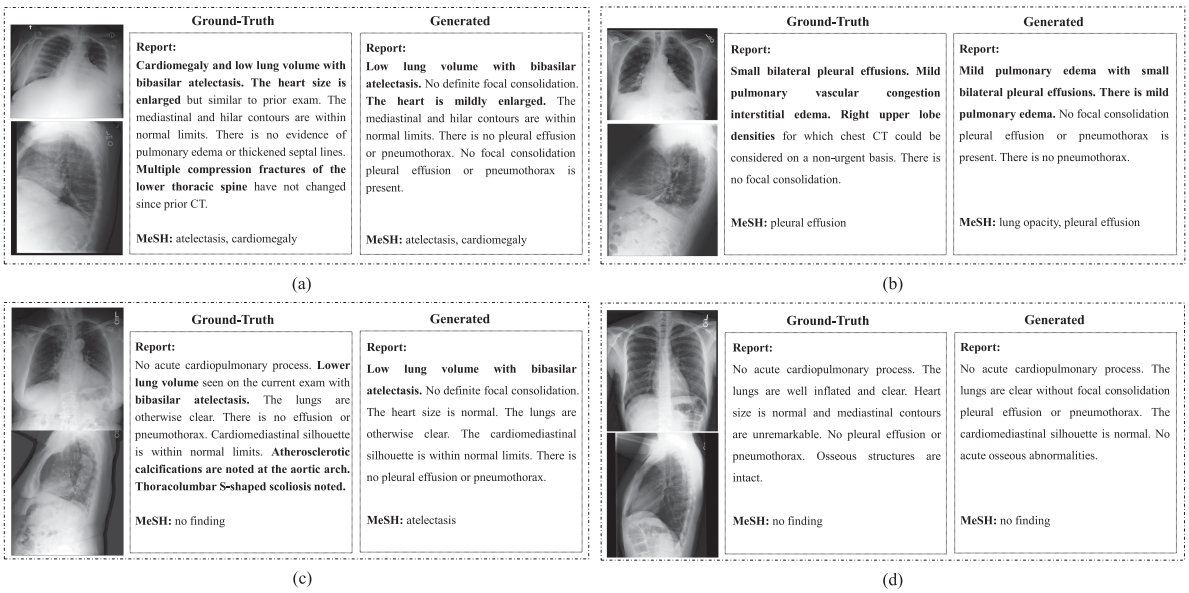


Fig. 5. Examples of generated reports given the frontal and lateral images in MIMIC-CXR. Highlighted words are descriptions for abnormality findings. (a), (b), and (c) are examples with abnormalities. (d) is the example without abnormality. The associated medical subject headings (MeSH) are presented at the bottom. The reference MeSH are generated by the automatic labeler CheXpert [38] from the ground-truth reports.

generated reports. Also, the long-tailed distribution of words in the vocabulary will affect the prediction. For example, the word frequency in IU X-Ray dataset is: "the" 3849, "no" 2705, "is" 2640, "are" 2638, "normal" 2152,..., "fossae" 1, "degenerated" 1. As can be observed, words like "the" and "no" appear much more frequently, while words like "fossae" and "degenerated" appear only once which however are usually used to describe the abnormal findings. Because the word generation is formulated as the multi-class classification at each time step (we output the word with the biggest probability), the model will tend to generate words that appeared frequently in the training set.

We can also observe that, in most cases, different sentences in the generated report will have different topics. But sometimes our model will repeat generating the same sentences as shown in (b) of Fig. 4: "No focal airspace opacity pleural effusion or pneumothorax. No focal airspace opacity pleural effusion or pneumothorax". Besides, a phenomenon should be noted that it can be hard to generate unpredictable sentences which compare

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

the current study with earlier historical studies, such as "Multiple compression fractures of the lower thoracic spine have not changed since prior CT" as shown in (a) of Fig. 5. Also, it can be hard to predict recommendations for further diagnoses, such as the "Right upper lobe densities for which chest CT could be considered on a non-urgent basis." as shown in (b) of Fig. 5.

For the MeSH prediction, as we can observe from Fig. 4 and Fig. 5, our method can generate MeSH accurately. In MIMIC-CXR dataset, as can be observed in (c) of Fig. 5, our method predicts the disease "atelectasis" accurately while the reference label "no finding" provided by the auto-labeler CheXpert is wrong. But sometimes some of the diseases may be missed (as shown in (a) and (b) of Fig. 4) or wrongly detected.

E. Quantitative Results and Comparison

In this section, we compare the performance of our method "Ours" with the following six MeIRG methods which also follow the encoder-decoder framework:

1) CNN-RNN [3]: First, multi-view visual features are extracted by CNN from medical images to initialize the RNN. Then, medical reports are generated with the RNN. For comparison, we extend the original work by replacing the LSTM with hierarchical LSTM.
2) Show, Attend and Tell (SAT) [5]: First, visual features are extracted by CNN and the soft-attention mechanism is adopted to obtain related visual features. Then, words are generated with LSTM. We extend the original work by replacing the LSTM with hierarchical LSTM.
3) Co-attention (CoAtt) [6]: First, visual features are extracted by CNN and semantic tags are predicted with the multi-label classification. Then, co-attention mechanism is employed to simultaneously attend to the visual features and the predicted tags. Also, a regularization term is used for the co-attention.
4) Text-Image Embedding Network (TieNet) [8]: First, visual features are extracted by CNN. Then, the report is generated word-by-word with an LSTM conditioned on the previous hidden state, previously generated words and CNN features. In addition, attention-encoded text embedding and saliency weighted global average pooling are employed for the disease classification.
5) Multimodal Recurrent Model with Attention (MRMA) [29]: First, the global visual feature is extracted by CNN and a start sentence is firstly generated with an LSTM, where the global visual feature is used as the initial input to predict the first word of the start sentence. Then, succeeding sentences are generated by taking the context of the previous sentence and the attended regional visual information as input to the LSTM, where the regional visual feature is used as the initial input of the LSTM to predict the first word of the succeeding sentence.
6) Multi-view encoder with hierarchical decoder, late fusion attention and medical concepts (MvH-AttL-MC) [34]: First, the image encoder is pre-trained with a large number of multi-view chest X-ray images. Then, multi-view visual features are extracted with sentence-level attention in a late fusion fashion. Besides, medical concepts are employed to assist report generation with word-level attention in the word decoder.

Among these methods, CoAtt, TieNet and MvH-AttL-MC also exploit semantic information to assist report generation. Nevertheless, they mostly exploit semantic tag information to assist report generation without exploiting report information. The performance comparisons on IU X-Ray and MIMIC-CXR are presented in Table III. Our method can generate reports with excellent performance compared with the recent works. Our model achieves the best performance across all metrics in IU X-Ray, improving BlEU-4 from 0.173 to 0.180, ROUGE-L from 0.383 to 0.398, CIDEr from 0.381 to 0.439. Also, our method achieves the best performance across metrics in MIMIC-CXR, improving BlEU-4 from 0.130 to 0.143, ROUGE-L from 0.316 to 0.326, CIDEr from 0.239 to 0.273.

Comparisons of the reports generated by "CNN-RNN," "Ours-visual-only" and "Ours" are presented in Fig. 6. As can be observed, "CNN-RNN" model depicts only normal findings and always repeats the same sentences. "Ours-visual-only" can describe findings better than "CNN-RNN," in which some disease descriptions can be generated. The phenomenon shows the attention mechanism can be helpful for report generation. "Ours" can depict most of the diseases and can generate more accurate descriptions compared with "Ours-visual-only". This phenomenon shows the semantic embeddings in our method are also helpful for the MeIRG, especially for the abnormal findings. We can conclude that both the attention mechanism and the semantic embeddings can be helpful for MeIRG.

F. Ablation Studies of Semantic Embeddings and Loss Terms

To verify the performance of the proposed semantic branches, ablation studies are also conducted. The ablation studies "Ours-visual-only," "Ours-no-MeRP" and "Ours-no-MeSH" in IU X-Ray and MIMIC-CXR are presented in Table IV. "Ours-visual-only" is to generate reports only with the attended visual embedding. "Ours-no-MeSH" is to generate reports with the attended visual embedding and MeRP embedding. "Ours-no-MeRP" is to generate reports with the visual attention embedding and the MeSH embedding. From the results, we can observe that the performance of "Ours-no-MeRP" and "Ours-no-MeSH" are better than the "Ours-visual-only". The performance of "Ours" is better than "Ours-visual-only," "Ours-no-MeRP" and "Ours-no-MeSH". We can conclude that the proposed semantic embeddings of the MeSH and MeRP are both helpful for the MeIRG. The main reason for the good performance is that semantic embedding branches will provide additional semantic features for MeIRG. On the one hand, MeSH embedding branch can predict disease labels such as "normal," "degenerative changes," "emphysema," "hyperexpansion," "atelectasis," "cardiomegaly," etc. It has also been experimented in [6][8] [34] that the semantic disease information can assist the MeIRG. On the other hand, MeRP embedding branch can predict additional disease attributes like the shape, location and size of the diseases such as "osseous", "tortuous", "atherosclerotic", "left-sided", etc. It is known that TF-IDF can extract keywords of the input report within the report corpus, similar reports will have similar TF-IDF [9]. In our method, we predict MeRP TF-IDF features to

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS IN IU X-RAY AND MIMIC-CXR. BLEU-N MEANS THE BLEU SCORES WITH N-GRAMS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
IU X-Ray	CNN-RNN [3]	0.351	0.252	0.187	0.142	0.366	0.291
	SAT [5]	0.391	0.264	0.190	0.143	0.323	0.273
	CoAtt [6]	0.434	0.307	0.227	0.173	0.383	0.381
	TieNet [8]	0.357	0.245	0.176	0.129	0.284	0.173
	MRMA [29]	0.367	0.260	0.196	0.147	0.337	0.209
	MvH-AttL-MC [33]	0.436	0.312	0.229	0.170	0.372	0.328
	Ours	**0.478**	**0.344**	**0.248**	**0.180**	**0.398**	**0.439**
MIMIC-CXR	CNN-RNN [3]	0.245	0.142	0.095	0.068	0.218	0.112
	SAT [5]	0.288	0.195	0.144	0.109	0.302	0.196
	CoAtt [6]	0.333	0.229	0.169	0.127	0.316	0.218
	TieNet [8]	0.279	0.173	0.113	0.076	0.210	0.118
	MRMA [29]	0.294	0.174	0.120	0.086	0.237	0.110
	MvH-AttL-MC [33]	0.330	0.227	0.170	0.130	0.315	0.239
	Ours	**0.362**	**0.251**	**0.188**	**0.143**	**0.326**	**0.273**

Ground-Truth

No acute intrathoracic process. PA and lateral views of the chest provided. **Lung volume are low.** There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.

CNN-RNN

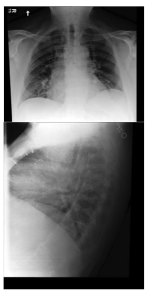
No acute cardiopulmonary process. The lungs are clear without focal consolidation. The lungs are clear without focal consolidation. The cardiomediastinal silhouette is normal. The lungs are clear. No pleural effusion or pneumothorax is seen.

Ours-visual-only

No acute cardiopulmonary process. PA and lateral views of the chest provided. The lungs are clear without focal consolidation. The lungs are clear without focal consolidation. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities.

Ours

No acute cardiopulmonary process. PA and lateral views of the chest provided. The lungs are clear without focal consolidation. The cardiomediastinal silhouette is normal. No pleural effusion or pneumothorax is seen. No free air below the right hemidiaphragm is seen.



Ground-Truth

Acute interstitial pulmonary edema. PA and lateral chest radiographs were obtained. **Lung volume are low. Bilateral basilar interstitial abnormality** is new. **Moderate cardiomegaly** is similar. There is no effusion consolidation or pneumothorax. Sternotomy wires are intact.

CNN-RNN

No acute cardiopulmonary process. The lungs are clear without focal consolidation. The lungs are clear without focal consolidation. The cardiomediastinal silhouette is normal. The cardiac and mediastinal silhouettes are unremarkable. No pleural effusion or pneumothorax is seen.

Ours-visual-only

Mild pulmonary edema. No focal consolidation. **The heart is mildly enlarged.** The mediastinal and hilar contours are unchanged. There is no pleural effusion or pneumothorax. No focal consolidation pleural effusion or pneumothorax is present

Ours

Increased interstitial markings bilaterally may be due to chronic interstitial lung disease. The heart is mildly enlarged. The heart is mildly enlarged. There is no pleural effusion or pneumothorax. There is no pleural effusion or pneumothorax.

Fig. 6. Comparison of the reports generated by "CNN-RNN," "Ours-visual-only" and "Ours". The highlighted words are descriptions of the abnormalities. The sample on the top is without abnormality. The sample at the bottom is with abnormalities.

TABLE IV
ABLATION STUDIES OF THE PROPOSED METHOD IN IU X-RAY AND MIMIC-CXR. BLEU-N MEANS THE BLEU SCORES WITH N-GRAMS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
IU X-Ray	Ours-visual-only	0.410	0.290	0.212	0.156	0.332	0.312
	Ours-no-MeRP	0.459	0.332	0.241	0.176	0.392	0.376
	Ours-no-MeSH	0.450	0.314	0.230	0.173	0.376	0.312
	Ours	**0.478**	**0.344**	**0.248**	**0.180**	**0.398**	**0.439**
MIMIC-CXR	Ours-visual-only	0.297	0.176	0.120	0.080	0.232	0.156
	Ours-no-MeRP	0.354	0.237	0.175	0.132	0.303	0.163
	Ours-no-MeSH	0.348	0.238	0.178	0.134	0.319	0.227
	Ours	**0.362**	**0.251**	**0.188**	**0.143**	**0.326**	**0.273**

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ABLATION STUDIES OF THE LOSS TERMS IN IU X-RAY AND MIMIC-CXR DATASETS. BLEU-N MEANS THE BLEU SCORES WITH N-GRAMS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
IU X-Ray	Ours-no-ℓ_{sem}	0.413	0.270	0.192	0.139	0.377	0.170
	Ours-no-ℓ_{stop}	0.289	0.201	0.150	0.113	0.390	0.201
	Ours-only-ℓ_{word}	0.253	0.162	0.114	0.084	0.351	0.196
	Ours	**0.478**	**0.344**	**0.248**	**0.180**	**0.398**	**0.439**
MIMIC-CXR	Ours-no-ℓ_{sem}	0.309	0.210	0.159	0.121	0.319	0.157
	Ours-no-ℓ_{stop}	0.275	0.184	0.139	0.107	0.294	0.126
	Ours-only-ℓ_{word}	0.229	0.143	0.103	0.075	0.238	0.066
	Ours	**0.362**	**0.251**	**0.188**	**0.143**	**0.326**	**0.273**

TABLE VI
EFFECTS OF MERP CORPUS CONSTRUCTION IN IU X-RAY AND MIMIC-CXR DATASETS. BLEU-N MEANS THE BLEU SCORES WITH N-GRAMS

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
IU X-Ray	Ours-200-MeRP	0.431	0.281	0.197	0.142	0.381	0.329
	Ours-1000-MeRP	0.454	0.305	0.220	0.162	**0.413**	0.330
	Ours	**0.478**	**0.344**	**0.248**	**0.180**	0.398	**0.439**
MIMIC-CXR	Ours-2000-MeRP	0.343	0.237	0.177	0.134	0.319	0.222
	Ours-10000-MeRP	0.359	0.248	0.185	0.142	0.321	0.234
	Ours	**0.362**	**0.251**	**0.188**	**0.143**	**0.326**	**0.273**

assist MeIRG by minimizing the distance between the predicted and the ground-truth TF-IDF features.

We also conduct ablation experiments to analyse the effects of three sub-loss terms in the loss function. Results are shown in Table V. "Ours-no-ℓ_{sem}" means to exclude semantic embedding loss in the loss function. "Ours-no-ℓ_{stop}" means to exclude stop control loss in the loss function. "Ours-only-ℓ_{word}" means only to include word generation loss in the loss function. As can be observed, the performance reduces largely when the ℓ_{stop} is excluded, since ℓ_{stop} will guide the topic encoding LSTM whether to encode a topic for the sentence decoding module. We can also observe that ℓ_{sem} will affect the report generation but the performance is still comparable with the model "Ours-visual-only" in Table IV when ℓ_{sem} is excluded. It is worth noting that the values of all evaluation metrics will be almost all zeros when ℓ_{word} is excluded.

G. Analysis of Time Efficiency, MeRP Corpus Construction and MeSH Prediction

We have conducted experiments to analyze the time efficiency of our method. For IU X-Ray, our model takes around 200 s on average to train an epoch and 9 s on average to test 300 studies. For MIMIC-CXR, our model takes around 45 minutes on average to train an epoch and 1.5 minutes on average to test 8000 studies. It is noteworthy that the TF-IDF features are previously constructed for training and the time is not included. The time efficiency of our method is superior to MRMA, similar to CoAtt, MvH-AttL-MC and TieNet, while inferior to the CNN-RNN and SAT. We also analyse the model parameters and the Giga Floating Point Operations (GFLOPs). The model is of 60.24 million parameters with 8.58 GFLOPs in IU X-Ray. The model is of 63.75 million parameters with 8.61 GFLOPs in MIMIC-CXR.

TABLE VII
PERFORMANCE COMPARISON OF MESH PREDICTION IN THE IU X-RAY. (ACC IS THE ACCURACY, PRE IS THE PRECISION, SPE IS THE SPECIFICITY)

Method	Average Acc	Average Spe	Average Pre	Average Recall
CoAtt [6]	0.935	0.955	0.160	0.177
TieNet [8]	0.938	**0.958**	0.187	0.165
Ours	**0.940**	0.956	**0.249**	**0.186**

As mentioned above, we obtained 2,339 and 62,801 reports from the training set as MeRP corpus for IU X-Ray and MIMIC-CXR respectively. To test the effects of MeRP corpus construction, we hereby construct another two MeRP corpus: randomly selecting 200 (Ours-200-MeRP) or 1000 reports (Ours-1000-MeRP) from the training set to construct MeRP corpus in IU X-Ray; randomly selecting 2000 (Ours-2000-MeRP) or 10000 reports (Ours-10000-MeRP) from the training set to construct MeRP corpus in MIMIC-CXR. Performance comparisons are shown in Table VI. As we can observe, the performance will reduce slightly when the corpus becomes smaller. The reason is that the larger the corpus is, the more accurate semantic information can be obtained.

For MeSH prediction, the performance comparisons in IU X-Ray and MIMIC-CXR are presented in Table VII and VIII. We employ the average accuracy, specificity, precision and recall across 14 disease labels and the top 30 MeSH to evaluate the performance in MIMIC-CXR and IU X-Ray respectively. As can be observed, "Ours" shows better performance than "CoAtt" and "TieNet" in the average accuracy and average precision. The recall of "Ours" is higher than "TieNet" but lower than "CoAtt"

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
PERFORMANCE COMPARISON OF MESH PREDICTION IN MIMIC-CXR. (ACC IS THE ACCURACY, PRE IS THE PRECISION, SPE IS THE SPECIFICITY)

Method	Average Acc	Average Spe	Average Pre	Average Recall
CoAtt [6]	0.916	0.915	0.254	**0.349**
TieNet [8]	0.921	**0.941**	0.204	0.286
Ours	**0.922**	0.919	**0.237**	0.326

in MIMIC-CXR. "TieNet" shows better performance in average specificity than "CoAtt" and "Ours". Overall, we can see that our approach achieves better performance in the average accuracy and average precision.

V. CONCLUSION

In this paper, we jointly encode deep visual and semantic embeddings through a triple-branch network (TriNet) for MeIRG. The MeRP embedding branch and MeSH embedding branch predict semantic embeddings of both the MeRP and the MeSH, while the visual attention branch captures visual embeddings from medical images with the soft-attention mechanism. Extensive experiments conducted on two benchmark datasets show that, with the assistance of semantic embeddings, our model can encode comprehensive representations for MeIRG and thus achieve good performance. As for our future work, we will focus on generating more accurate and fine-grained reports, especially for abnormality findings. Also, we will try to investigate how medical report corpus can be further exploited as the prior knowledge for medical imaging report generation, such as utilizing a pre-trained language model and constructing the knowledge graph.

REFERENCES

[1] P. Kisilev *et al.*, "From medical image to automatic medical report generation," *IBM J. Res. Dev.*, vol. 59, no. 2-3, pp. 1–7, Mar. 2015.

[2] H.-C. Shin *et al.*, "Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation," *J. Mach. Learn. Res.*, vol. 17, no. 107, pp. 1–31, 2016.

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

[4] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 317–325.

[5] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[6] B. Jing, P. Xie, and E. P. Xing, "On the automatic generation of medical imaging reports," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2577–2586.

[7] Y. Zhang *et al.*, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12 910–12917.

[8] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9049–9058.

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] M. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019.

[12] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.

[13] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.

[14] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013.

[15] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[16] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2743–2754, Jun. 2019.

[17] J. H. Tan, C. S. Chan, and J. H. Chuah, "Comic: Toward a compact image captioning model with attention," *IEEE Trans. Multimedia*, vol. 21, pp. 2686–2696.

[18] N. Xu *et al.*, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Trans. Multimedia*, vol. 22, pp. 1372–1383.

[19] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, May 2020.

[20] L. Guo, J. Liu, S. Lu, and H. Lu, "Show, tell, and polish: Ruminant decoding for image captioning," *IEEE Trans. Multimedia*, vol. 22, pp. 2149–2162.

[21] X. Li *et al.*, "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Trans. Multimedia*, vol. 21, pp. 2347–2360.

[22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2016, pp. 4651–4659.

[23] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1151–1159.

[24] B. Dai, D. Lin, R. Urtasun, and S. Fidler, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2989–2998.

[25] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4155–4164.

[26] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2020, pp. 9962–9971.

[27] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 685–10694.

[28] H.-C. Shin *et al.*, "Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2497–2506.

[29] Y. Xue *et al.*, "Multimodal recurrent model with attention for automated radiology report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2018, pp. 457–466.

[30] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Adv. Neural Inf. Process. Syst.*, 2018, pp. 1537–1547.

[31] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6666–6673.

[32] B. Jing, Z. Wang, and E. P. Xing, "Show, describe and conclude: On exploiting the structure information of chest X-ray reports," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6570–6580.

[33] P. Harzig, Y.-Y. Chen, C. F., and R. Lienhart, "Addressing data bias problems for chest X-ray image report generation," in *Proc. Brit. Mach. Vis. Conf.*, no. 144, 2019, pp. 1–11.

[34] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2019, pp. 721–729.

[35] X. Xie *et al.*, "Attention-based abnormal-aware fusion network for radiology report generation," in *Database Syst. Adv. Appl.*, 2019, pp. 448–452.

[36] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2015.

[37] A. E. Johnson *et al.*, "Mimic-CXR, a de-identifed publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 317, 2019.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.

[38] J. Irvin *et al.*, Seekins, and others "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 590–597.
[39] Y. Peng *et al.*, "NegBio: A high-performance tool for negation and uncertainty detection in radiology reports," in *AMIA Inform. Summit*, 2018, pp. 188–196.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[42] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[43] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[44] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out, Post- Conf. Workshop ACL*, 2004, pp. 74–81.

Yan Yang (Student Member, IEEE) received the B.Eng. and M.Sc. degrees from Ningbo University, Zhejiang, China. She is currently working toward the Ph.D. degree with the Key Laboratory of Complex Systems Modelling and Simulation in the School of Computer Science, Hangzhou Dianzi University, Hangzhou, China. Her research interests include machine learning, vision and language, and medical image analysis.

Jun Yu (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 include multimedia analysis, machine learning, and image processing. In 2017, he was the recipient of the IEEE SPS Best Paper Award. Dr. Yu has cochaired several special sessions, invited sessions, and workshops. He served as a program committee member or reviewer of top conferences and prestigious journals. He is a Professional Member of the Association for Computing Machinery and China Computer Federation.

Jian Zhang received the Ph.D. degree from Zhejiang University, Zhejiang, China. He is currently a Professor with the School of Science and Technology, Zhejiang International Studies University, Hangzhou, China. From 2009 to 2011, he was a Postdoctoral Research Fellow with the Department of Mathematics, Zhejiang University. In 2016, he had been doing research on machine learning with Simon Fraser University, as a Visiting Scholar. His research interests include but are not limited to machine learning and image processing. He served as a Reviewer of several prestigious journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Weidong Han received the bachelor's degree and the Ph.D. degree from College of Medicine, Zhejiang University, Zhejiang, China. He is currently a Physician and with the Department of Medical Oncology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China. He has authored or coauthored more than 80 articles with more than 3000 citations. His research interests include but are not limited to application of imaging omics in tumor diagnosis and treatment. He served as a Reviewer of several prestigious journals, including the *Advanced Science*, *Biomaterials*, *Signal Transduction and Targeted Therapy*, and *Theranostics*.

Hanliang Jiang received the bachelor's degree in internal medicine and the Ph.D. degree in basic medicine from Zhejiang University, Zhejiang, China.

He is studying at lung cancer mechanism and comprehensive therapy. He is currently a Physician and with the Pulmonary and Critical Care Medicine, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China.

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.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on August 27,2023 at 01:47:21 UTC from IEEE Xplore. Restrictions apply.