Adaptive Latent Graph Representation Learning for Image-Text Matching

Mengxiao Tian, *Student Member, IEEE*, Xinxiao Wu, *Member, IEEE*, and Yunde Jia, *Member, IEEE*

***Abstract*— Image-text matching is a challenging task due to the modality gap. Many recent methods focus on modeling entity relationships to learn a common embedding space of image and text. However, these methods suffer from distractions of entity relationships such as irrelevant visual regions in an image and noisy textual words in a text. In this paper, we propose an adaptive latent graph representation learning method to reduce the distractions of entity relationships for image-text matching. Specifically, we use an improved graph variational autoencoder to separate the distracting factors and latent factor of relationships and jointly learn latent textual graph representations, latent visual graph representations, and a visual-textual graph embedding space. We also introduce an adaptive cross-attention mechanism to perform feature attending on the latent graph representations across images and texts, thus further narrowing the modality gap to boost the matching performance. Extensive experiments on two public datasets, Flickr30K and COCO, show the effectiveness of our method.**

***Index Terms*— Image-text matching, latent representation learning, graph variational autoencoder.**

I. Introduction

IMAGE-TEXT matching has achieved remarkable progress in a variety of applications, such as cross-modal retrieval [1], [2], image captioning [3], [4], and visual question answering [5], [6]. Image-text matching is a challenging task due to the fact that there exists a large modality gap between image and text. Many methods [7], [8] have successfully used deep neural networks to extract the global visual features of an image and the global features of a text, respectively, and then map these two global features into a common embedding space for similarity evaluation. Some local representation based methods [9], [10] seek to align local visual regions and textual words, which effectively capture the similarities between different modalities. However, they often overlook the rich entity relationships contained in images and texts, which has been demonstrated to play pivotal roles in video understanding [11], [12] and text classification [13], [14].

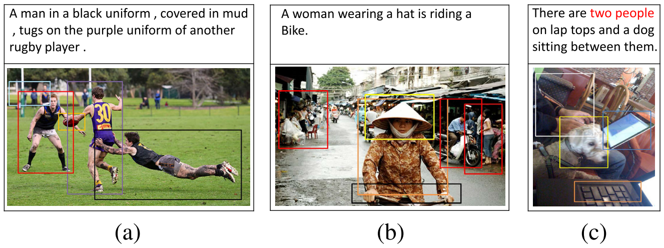

A man in a black uniform , covered in mud , tugs on the purple uniform of another rugby player .
A woman wearing a hat is riding a Bike.
There are two people on lap tops and a dog sitting between them.
(a)
(b)
(c)

Fig. 1. Examples of prominent distracting factors of relationships between entities in both image and text: (a) distraction from the irrelevant object region in the red bounding box, (b) distraction from the irrelevant background in the red bounding boxes, and (c) distraction from noisy words in red font.

More recent methods [15], [16], [17], [18] use graph structures to capture the relationships between entities in both image and text for aligning image and text, which have achieved good performance. However, these methods suffer from the distracting factors of entity relationships in both image and text, such as irrelevant objects and background regions in images, and noisy words and phrases in texts, leading to the false cross-modal alignments. Fig. 1(a) shows an example of irrelevant object regions in an image. The image contains salient objects of “man in black uniform”, “man in purple uniform” and “rugby”, whereas the corresponding text “A man in a black uniform, covered in mud, tugs on the purple uniform of another rugby player” does not have the relevant description of the onlooker in purple uniform. Fig. 1(b) shows an example of irrelevant background regions in an image. The image has the background regions of “man riding a motorcycle”, “woman in pink”, and “some people who walk”, which does not correspond to any description in the text “A woman wearing a hat is riding a bike”. The aforementioned irrelevant objects and background regions in images may result in misalignments between visual regions and textual words. Fig. 1(c) demonstrates an example of noisy words or phrases in the text, where the phrase “two people” is invisible in the image.

To address the issue of distracting factors, in this paper, we propose an adaptive latent graph representation learning

Manuscript received 19 February 2022; revised 16 September 2022 and 21 November 2022; accepted 4 December 2022. Date of publication 19 December 2022; date of current version 30 December 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62072041 and Grant 62176021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ming Shao. *(Corresponding author: Xinxiao Wu.)*

Mengxiao Tian is with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: mengxiao_tian@bit.edu.cn).

Xinxiao Wu is with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China, and also with the Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: wuxinxiao@bit.edu.cn).

Yunde Jia is with the Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China, and also with Beijing Key Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: jiayunde@smbu.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3229631

method that reduces the distracting factors of entity relationships for accurate image-text matching. Our method is inspired by variational autoencoder [19] that disentangles modality-specific factors of observational data through KL divergence between the posterior distribution and a standard Gaussian prior distribution. Specifically, we first build a visual graph for an image and a textual graph for a sentence.

A node in the visual graph represents a visual object, and an edge represents a visual relationship between two connected objects; a node in the textual graph represents a word, and an edge represents a semantic relationship between two words. Then we introduce an improved graph variational autoencoder to jointly learn latent textual graph representation, latent visual graph representation, and a common graph embedding space shared by image and text. Benefiting from the ability of variational autoencoder to partition the underlying explanatory factors in latent space, the autoencoder separates the alignable latent factor and the unalignable distracting factors. This is done by using a variational lower bound to model the approximate posterior distribution of the visual or the textual graph representation. By filtering out the distracting factors, the alignable latent factor can be regarded as the latent modality-invariant graph representation, narrowing the gap between visual and textual modalities. The learned latent graph representations of image and text are projected into a common latent space where the distributions of the two graph representations are aligned by minimizing the Wasserstein distance between them.

To further narrow the gap between visual and textual modalities, we introduce an adaptive cross-attention mechanism to perform feature attention on the latent graph representations of image and text. The attention weights assigned to the latent representations of different modalities are adaptively adjusted to facilitate concentrating on more salient and important entity relationships for image-text matching. We also introduce a two-phase training strategy to train the graph construction network and then optimize the whole network to improve the discriminative ability of the disentangled representations.

The remainder of this paper is organized as follows. In Section II, we summarize previous works related to our method. Section III describes an adaptive latent graph representation learning method for image-text matching. Section IV discusses experimental results on two public datasets, and conclusion is given in Section V.

II. Related Work

A. Image-Text Matching

From the perspective of feature representation learning, the methods of image-text matching can be roughly divided into two categories: global representation learning and local representation learning. The global representation learning methods [20], [21] employ pre-trained deep neural networks to extract the global features of images and texts, and learn a common embedding space between the visual and textual features. Frome et al. [21] proposed a visual-semantic embedding model that extracts visual features of images by CNN and textual features of texts by Skip-Gram. In recent years, most approaches [22], [23] employ varieties of RNN architecture to capture the long-range contextual information of language. Kiros et al. [22] used LSTM to extract the global semantic representations of texts. Faghri et al. [23] also used the global features of images and texts extracted by CNN and GRU, respectively, and designed a novel objective function with hard negative mining to further improve the matching accuracy.

The local representation based methods [24], [25] extract local patterns to align visual regions and words for image-text matching. Karpathy et al. [24] detected visual regions in images by the pre-trained R-CNN, and then learned the similarities between words in sentences and regions in images. Motivated by the success of bottom-up attention mechanism [26], Lee et al. [27] presented a stacked cross-attention model to aggregate the local similarity matching results between image regions and words. Wang et al. [28] proposed a cross-modal adaptive message passing method to learn interactions across regions and words. To capture the contextual information of different modalities, Qu et al. [29] proposed a gating self-attention for context modeling to enhance image and text representations. Later, they designs four connected network cells to capture both inter-modal interactions and intra-modal relationships [30].

In order to reduce some noises which are often uncorrelated to other modalities, several methods have been proposed by introducing various loss functions [31], [32]. Song et al. [31] employed a diversity loss to penalize the redundancy for learning diverse embedding representations of an instance. Huang et al. [32] randomly shuffled the captions of training images as noise in different proportions, and rectified the noise by using an adaptive prediction function and a novel triplet loss. In contrast, our method focuses on learning well-separated representations by filtering out the irrelevant information in each modality. Moreover, our method uses graph structures to capture the relationships between entities and learns a common graph embedding space shared of image and text.

B. Graph Representation Learning

The graph representation has been commonly used to model relationships between entities in vision-language tasks, including image captioning [33], [34], [35], visual question answering [36], [37] and visual common sense reasoning [38], [39]. Some image-text matching methods [7], [15], [16], [40] employ graph structures to enhance the monolithic representations of visual and textual contents, and achieve a good performance. Li et al. [7] proposed a visual semantic reasoning method that learns relationships between object regions in images to generate enhanced visual representations. Liu et al. [15] proposed a graph matching network to perform node-level and structure-level matching between images and texts. Wang et al. [16] built scene graphs to represent images and texts, and then performed object-level and relationship-level matching between images and texts on the scene graphs. However, these methods suffer from the prominent distracting factors of relationships. To address this problem, we introduce the variational autoencoder into the

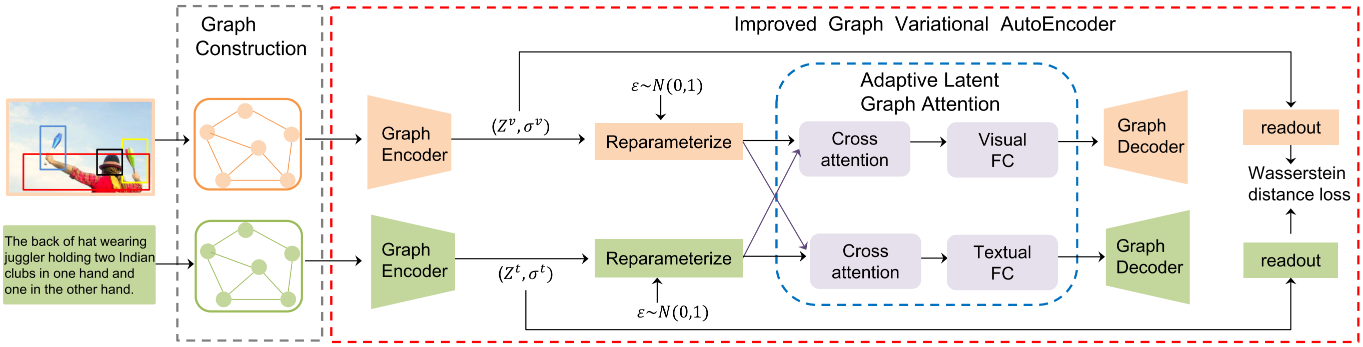


Fig. 2. Overview of the adaptive latent graph representation method.

graph representation learning to separate the distracting factors and the latent factor of entity relationships. Reference [40] is more related to our work, which calculates the distraction scores by using an information-theoretic entropy to quantify visual distractions. Our work mainly focuses on exploring a disentanglement learning method to filter out the distracting factors for image-text matching task.

C. Latent Representation Learning

Latent representation learning aims to identify and separate the explanatory features that are relatively unaffected by other feature changes, and has been widely studied in few-shot learning learning [41], [42], zero-shot learning [43], [44], [45], [46], and cross-modal retrieval [47], [48], [49], [50], [51]. The most relevant methods of latent representation learning to our work are [44] and [49]. Ye and Shen [44] presented a graph metric learning method that can learn graph representations by introducing variational structures without taking into account the relationship information between nodes for graph encoding. Unlike their work, our method adaptively learns the importance of neighboring nodes in the encoding process to enhance the embedding ability of the graph encoder. Fu et al. [49] presented a stochastic latent variable model to achieve the alignment between recipes and food images by extracting the global feature vectors of images and texts. Different from their method, our method introduces graph structure into variational autoencoder to capture the fine-grained relationships between entities for image-text matching in a shared latent graph representation space.

III. Our Method

Our method consists of two modules: a graph construction module and an improved graph variational autoencoder. The graph construction module constructs a visual graph and a textual graph, respectively, to represent the entity relationships in image and text. The improved graph variational autoencoder learns a common latent representation space of the visual and textual graphs via separating the distracting factors and the latent factor of relationships. Then the latent representations of the visual and textual graphs get attention by using the adaptive latent graph attention to focus on more salient and important entity relationships. Fig. 2 shows the overview of our method, where the improved graph variational autoencoder is our main contribution.

To train our model, we introduce a two-phase training strategy, where the graph construction module is pre-trained using the ranking loss $\mathcal{L}_{rank}$ and then the whole network is trained using the whole loss $\mathcal{L}$.

A. Graph Construction

1) Visual Graph: We construct a visual graph $G_v = (V_v, E_v)$ to model the structural information among visual objects in an image. Each node $v \in V_v$ denotes a visual object and each edge $e \in E_v$ denotes the visual relation between the two connected objects. We use a region-level feature extracted by an image embedding module [29] to represent a node, and we then perform self-attention over the region-level feature. The edge is initialized by heuristic information of the semantic similarity between the two connected objects. Denote $\boldsymbol{A}^v \in \mathbb{R}^{|V_v|\times|V_v|}$ as the adjacent matrix of nodes, where $|V_v|$ is the number of nodes. If there exists an edge from node v_i to node v_j, we set $\boldsymbol{A}^v(i,j) = 1$. In practice, we construct an undirected fully-connected graph to represent the semantic interactions between visual objects, and all the elements of $\boldsymbol{A}^v$ are set to 1. Subsequently, we normalize each row of $\boldsymbol{A}^v$ to make the sum of edge values connected to node v_i as 1. Let $\boldsymbol{X}^v \in \mathbb{R}^{|V_v|\times d}$ be the node features of a visual graph, where the i-th row $\boldsymbol{x}_i^v$ is the visual feature of node v_i, and $\boldsymbol{S}^v \in \mathbb{R}^{|V_v|\times|V_v|}$ be the similarity matrix of the nodes, where the element $s_{i,j}^v$ is the similarity between node v_i and node v_j. The edge weights of a visual graph are given by

$$\begin{aligned} \boldsymbol{W}_e^v &= \boldsymbol{S}^v \odot \boldsymbol{A}^v, \\ s_{ij}^v &= \frac{\exp(\mathrm{LeakyRelu}((\boldsymbol{W}_q^v\boldsymbol{x}_i^v)^\top(\boldsymbol{W}_k^v\boldsymbol{x}_j^v)))}{\sum_{j=0}^{|V_v|}\exp(\mathrm{LeakyRelu}((\boldsymbol{W}_q^v\boldsymbol{x}_i^v)^\top(\boldsymbol{W}_k^v\boldsymbol{x}_j^v)))}, \end{aligned} \quad (1)$$

where $\odot$ is element-wise multiplication, and $\boldsymbol{W}_q^v$ and $\boldsymbol{W}_k^v$ are the trainable parameters.

2) Textual Graph: We use an undirected fully-connected graph $G_t = (V_t, E_t)$ for each text. And each node $t \in V_t$ denotes a word, represented by a word embedding feature [29], and each edge $e \in E_t$ denotes the relationship between the two connected words, preliminary estimated by the semantic similarity between them. The topological structure of the textual graph G_t is represented by the adjacency matrix

$\boldsymbol{A}^t \in \mathbb{R}^{|V_t|\times|V_t|}$, where $|V_t|$ is the number of nodes. The undirected edge value is set to 1 if there exists an edge from node t_i to node t_j, and all the elements of the adjacent matrix are set to 1, i.e., $\boldsymbol{A}^t(i, j) = 1$. With regard to normalization, we perform a normalization of $\boldsymbol{A}^t$, similar to the normalization of $\boldsymbol{A}^v$. Let $\boldsymbol{X}^t \in \mathbb{R}^{|V_t|\times d}$ be the node features of the textual graph, where the i-th row $\boldsymbol{x}_i^t$ is the textual feature of node t_i, and $\boldsymbol{S}^t \in \mathbb{R}^{|V_t|\times|V_t|}$ be the similarity matrix, where the element $s_{i,j}^t$ is the similarity information between node t_i and node t_j. Similar to the calculation of the edge weights $\boldsymbol{W}_e^v$ of the visual graph, we calculate the edge weights of the textual graph as follows:

$$\begin{aligned}\boldsymbol{W}_e^t &= \boldsymbol{S}^t \odot \boldsymbol{A}^t,\\ s_{ij}^t &= \frac{\exp(\text{LeakyRelu}((\boldsymbol{W}_q^t\boldsymbol{x}_i^t)^\top(\boldsymbol{W}_k^t\boldsymbol{x}_j^t)))}{\sum_{j=0}^{|V_t|}\exp(\text{LeakyRelu}((\boldsymbol{W}_q^t\boldsymbol{x}_i^t)^\top(\boldsymbol{W}_k^t\boldsymbol{x}_j^t)))},\end{aligned} \tag{2}$$

where $\boldsymbol{W}_q^t$ and $\boldsymbol{W}_k^t$ are the learnable parameters.

B. Improved Graph Variational AutoEncoder

Given the constructed visual graphs of images and textual graphs of texts, the core issue resides in how to learn a common graph representation between the images and texts by capturing the entity relationships for image-text matching. To separate the distracting factors of entity relationships, we introduce an improved graph variational autoencoder that uses the variational autoencoder into graph encoding to learn the latent representations of visual and textual graphs.

1) Variational Autoencoder: The variational autoencoder is capable of disentangling the explanatory features in a latent space [52] by filtering out the modality-specific features of variations from observed data. It typically learns two kinds of representations for each modality. One is modality-invariant representation for reducing the modality gap between multimodal signals from different sources. The other is modality-specific representation for holding distinctive information for each modality. To achieve the modality-disentangled representation learning, the variational autoencoder employs a regularization of the KL divergence between the approximate posteriors and the priors in the latent representation space, and the priors impose modality-invariant features to be independent of the modality-specific features.

In concrete, the encoder in variational autoencoder converts the observed input data $\boldsymbol{c}$ into a latent conditional distribution, modeled by Gaussian distribution with the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. Then, a latent representation z is randomly sampled from the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Finally, the latent representation is used to reconstruct the input data by the decoder. The objective function of the variational autoencoder is defined as the variational lower bound to the log likelihood:

$$\mathbb{E}_{z\sim q_\theta(z|\boldsymbol{c})}[\log p_\phi(\boldsymbol{c}|z)] - \text{D}_{\text{KL}}(q_\theta(z|\boldsymbol{c})||p(z)), \tag{3}$$

where $p(z)$ is the prior distribution for z. $p_\phi(\boldsymbol{c}|z)$ can be treated as a decoder, and $q_\theta(z|\boldsymbol{c})$ can be treated as an encoder that is modeled by a factorized Gaussian distribution with a diagonal covariance matrix. The first term of Eq. 3 is a reconstruction error to observe how effectively the decoder learns to reconstruct $\boldsymbol{c}$ given its latent representation z. The second term is KL divergence to measure the distance between $q_\theta(z|\boldsymbol{c})$ and $p(z)$. Let's make it clear that a latent representation z is generated from the Gaussian distribution by using the reparametrization trick [53].

2) Graph Encoding: Different from [44] and [45] that both treat different neighbor nodes equally during the aggregation process by using a two-layer graph convolution network (GCN), we design a self-attention graph neural network (GNN) to assign different weights to different neighborhoods by taking into account the different contributions of neighborhoods to improve the graph encoding performance. Specifically, the edge weights are integrated into the self-attention graph layer for more flexibility, and the node features are updated through the self-attention mechanism

$$\boldsymbol{Z} = \text{ELU}(\boldsymbol{W}_e\boldsymbol{X}\boldsymbol{W}), \tag{4}$$

where $\boldsymbol{Z} \in \mathbb{R}^{|V|\times d}$, $\boldsymbol{W}_e$ is the normalized edge weight matrix of visual graph or textual graph described in Section III-A earlier, $\boldsymbol{X}$ is the input node features, $\boldsymbol{W}$ is the learnable parameter, and $\text{ELU}(\cdot)$ is an activation function.

For the visual graph, we design the graph encoder with a two-layer self-attention GNN to generate the updated node features Z^v, which is the output of the first self-attention GNN layer. For the textual graph, we combine a self-attention GNN and a gated graph neural network (GGNN) into a graph encoder. Given the input textual node features $\boldsymbol{X}^t$, we feed $\boldsymbol{X}^t$ into the self-attention GNN layer to generate the updated node features $\boldsymbol{Z}_1^t$, and then feed $\boldsymbol{Z}_1^t$ into the GGNN layer. We stack the GGNN layer m times to ensure that each node receives more information from its high-order neighbors. That is to say, the message from one node is propagated to another node in an m-hop way:

$$\begin{aligned}{\boldsymbol{Z}_1^t}^{(m)} &= {\tilde{\boldsymbol{Z}}_1^t}{}^{(m)} \odot \boldsymbol{U}^{(m)} + {\boldsymbol{Z}_1^t}^{(m-1)} \odot (1 - \boldsymbol{U}^{(m)}),\\ {\tilde{\boldsymbol{Z}}_1^t}{}^{(m)} &= \tanh(\boldsymbol{W}_{\text{z}}\boldsymbol{a}^{(m)} + \boldsymbol{F}_{\text{z}}(\boldsymbol{r}^{(m)} \odot {\boldsymbol{Z}_1^t}^{(m-1)}) + \boldsymbol{b}_{\text{z}}),\\ \boldsymbol{r}^{(m)} &= \text{sigmoid}(\boldsymbol{W}_r\boldsymbol{a}^{(m)} + \boldsymbol{F}_r{\boldsymbol{Z}_1^t}^{(m-1)} + \boldsymbol{b}_r),\\ \boldsymbol{U}^{(m)} &= \text{sigmoid}(\boldsymbol{W}_U\boldsymbol{a}^{(m)} + \boldsymbol{F}_U{\boldsymbol{Z}_1^t}^{(m-1)} + \boldsymbol{b}_u),\\ \boldsymbol{a}^{(m)} &= \boldsymbol{A}^t{\boldsymbol{Z}_1^t}^{(m-1)}\boldsymbol{W}_a,\end{aligned} \tag{5}$$

where $\boldsymbol{A}^t$ is the adjacency matrix of the textual graph, ${\boldsymbol{Z}_1^t}^{(m)}$ is the final updated node features of the textual graph, $\boldsymbol{W}_*$, $\boldsymbol{F}_*$ and $\boldsymbol{b}_*$ are the learnable parameters, and $\boldsymbol{U}^{(m)}$ and $\boldsymbol{r}^{(m)}$ are the update gate and reset gate, respectively.

3) Latent Graph Representation Learning: Our goal is to learn a common latent graph representation of image and text, which is shared by different modalities by discarding the modality-specific features to achieve the representation disentanglement. Based on the aforementioned graph encoding, the learned node features $\boldsymbol{Z}^v$ in the visual graph are used as the latent graph representation of the image, and the learned node features ${\boldsymbol{Z}_1^t}^{(m)}$ in the textual graph are used as the latent graph representation of the text.

Since the operation of sampling from the Gaussian distrbution is non-differentiable, we reparametrieze latent representation to make the gradient descent possible. Taking the visual graph for instance, the reparametrized latent representation is randomly sampled as follows:

$$\hat{\boldsymbol{Z}}^v = \boldsymbol{Z}^v + \boldsymbol{\sigma}^v \odot \epsilon^v, \tag{6}$$

where $\boldsymbol{\sigma}^v$ is the output of the second self-attention GNN layer and is used to model the modality-specific information for images, $\odot$ is the element-wise product, and $\epsilon^v \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an auxiliary noise variable. Notably, the prior distribution $p(\hat{\boldsymbol{Z}}^v)$ is assumed to be a Gaussian distribution that satisfies $p(\hat{\boldsymbol{Z}}^v) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, the approximation of posterior distribution $\hat{\boldsymbol{Z}}^v$ is represented as

$$q(\hat{\boldsymbol{Z}}^v|G_v) \sim \mathcal{N}(\boldsymbol{Z}^v, \boldsymbol{\sigma}^{v2}), \tag{7}$$

To enhance the representation ability of node features, the node features are reconstructed by a visual graph decoder $\mathrm{D_V}(\cdot)$, which is implemented by a linear layer with a sigmoid layer. Here, we design a smooth variational reconstruction loss by the L1 loss, which has little change in the gradient of the predicted value to ensure training stability. Benefiting from the variational reconstruction, the visual node features generated by the graph construction module contain informative and necessary details, and the approximate posterior estimation is more accurate, thus improving the discriminativeness and robustness of the latent graph representation. The reconstruction loss for the visual graph is given by

$$\mathcal{L}_r^v = \begin{cases} 0.5(\boldsymbol{X}^v - \boldsymbol{X}^{v\prime})^\top(\boldsymbol{X}^v - \boldsymbol{X}^{v\prime}), & |\boldsymbol{X}^v - \boldsymbol{X}^{v\prime}| < 1 \\ |\boldsymbol{X}^v - \boldsymbol{X}^{v\prime}| - 0.5, & otherwise \end{cases} \tag{8}$$

where $\boldsymbol{X}^v$ is the input visual node features, and $\boldsymbol{X}^{v\prime} = \mathrm{D_V}(\widetilde{\boldsymbol{Z}}^v)$ is the reconstructed node features via the visual graph decoder $\mathrm{D_V}(\cdot)$. $\widetilde{\boldsymbol{Z}}^v$ is the updated reparametrized latent representation via an adaptive latent graph attention module that will be illustrated in Section III-B.4 later.

Similarly, the error reconstruction function for the text graph is formulated as

$$\mathcal{L}_r^t = \begin{cases} 0.5(\boldsymbol{X}^t - \boldsymbol{X}^{t\prime})^\top(\boldsymbol{X}^t - \boldsymbol{X}^{t\prime}), & |\boldsymbol{X}^t - \boldsymbol{X}^{t\prime}| < 1 \\ |\boldsymbol{X}^t - \boldsymbol{X}^{t\prime}| - 0.5, & otherwise \end{cases} \tag{9}$$

Formally, the final loss of the improved graph variational autoencoder is defined as

$$\mathcal{L}_g = \mathcal{L}_r^v + \mathcal{L}_r^t + \mathcal{L}_{KL}, \tag{10}$$

where $\mathcal{L}_r^v$ and $\mathcal{L}_r^t$ represent the visual graph reconstruction loss and the textual graph construction loss, respectively. $\mathcal{L}_{KL}$ is the summation of KL divergences between the prior distribution and the posterior approximation distribution on both image and text:

$$\mathcal{L}_{KL} = D_{KL}(q(\hat{\boldsymbol{Z}}^v|G_v)||p(\hat{\boldsymbol{Z}}^v)) + D_{KL}(q(\hat{\boldsymbol{Z}}^t|G_t)||p(\hat{\boldsymbol{Z}}^t)) \tag{11}$$

The training objective of the improved graph variational autoencoder is to disentangle the common semantic information from the modality-specific information to avoid the distractions of the unalignable features.

The final latent graph representations for image and text are given by

$$\begin{aligned} \boldsymbol{e}^v &= \mathrm{readout}(\boldsymbol{Z}^v), \\ \boldsymbol{e}^t &= \mathrm{readout}(\boldsymbol{Z}_1^{t\,(m)}), \end{aligned} \tag{12}$$

where readout$(\cdot)$ denotes an average operation over the input features. Furthermore, to align the distributions of latent semantic representations towards modalities, we push the prior graph embedding distributions of image and text to be close by minimizing their Wasserstein distance loss

$$\mathcal{L}_{WD} = (||\boldsymbol{e}^v - \boldsymbol{e}^t||^2 + ||\mathrm{readout}(\boldsymbol{\sigma}^{\mathrm{v}}) - \mathrm{readout}(\boldsymbol{\sigma}^{\mathrm{t}})||^2)^{\frac{1}{2}}. \tag{13}$$

4) Adaptive Latent Graph Attention: To further narrow the modality gap between image and text, we use an adaptive cross-attention mechanism to capture the latent interactions between visual and textual representations. With the adaptive latent graph attention, our method focuses on the more common and important entity relationships in image and text by adjusting the attention weights across the latent representations of different modalities. The adaptive latent graph attention aggregates the local features of one modality using the cross-modal attention weights, and integrates the aggregated feature into the local features of the other modality, as shown in Fig 2.

Given the reparametrized latent visual representation $\hat{\boldsymbol{Z}}^v$ and the reparametrized latent textual representation $\hat{\boldsymbol{Z}}^t$, the attention weight is calculated by multiplication of $\hat{\boldsymbol{Z}}^v$ and $\hat{\boldsymbol{Z}}^t$, and then the output probability is produced by a softmax normalization function. Further, we aggregate the reparametrized latent representation of visual graph (or textual graph) with the attention weights and make concatenation with the opposite reparametrized latent representation to obtain the updated reparametrized latent representation of textual graph $\widetilde{\boldsymbol{Z}}^t$ (or updated reparametrized latent representation of visual graph $\widetilde{\boldsymbol{Z}}^v$). The whole process can be formulated as

$$\begin{aligned} \widetilde{\boldsymbol{Z}}^v &= \mathrm{FC}(\mathrm{concat}((\mathrm{softmax}(\hat{\boldsymbol{Z}}^v \hat{\boldsymbol{Z}}^{t\top}) \odot \hat{\boldsymbol{Z}}^t), \hat{\boldsymbol{Z}}^v)), \\ \widetilde{\boldsymbol{Z}}^t &= \mathrm{FC}(\mathrm{concat}((\mathrm{softmax}(\hat{\boldsymbol{Z}}^t \hat{\boldsymbol{Z}}^{v\top}) \odot \hat{\boldsymbol{Z}}^v), \hat{\boldsymbol{Z}}^t)), \end{aligned} \tag{14}$$

where concat$(\cdot, \cdot)$ is the concatenation operation, and FC$(\cdot)$ is the fully-connected layer that converts the enhanced latent representation into a h-dimensional feature vector.

C. Joint Training

To encourage the similarity scores of the matched image-text pairs to be larger than that of the mismatched pairs, we define the ranking loss as

$$\begin{aligned} \mathcal{L}_{rank} = &\max(0, m + S(\boldsymbol{e}^v, \boldsymbol{e}^{t-}) - S(\boldsymbol{e}^v, \boldsymbol{e}^t)) \\ &+\max(0, m + S(\boldsymbol{e}^{v-}, \boldsymbol{e}^t) - S(\boldsymbol{e}^v, \boldsymbol{e}^t)), \end{aligned} \tag{15}$$

where m is a margin factor, $S(\cdot)$ is the cosine similarity between two latent representations, $(\boldsymbol{e}^v, \boldsymbol{e}^t)$ is the matched image-text pair, and $(\boldsymbol{e}^v, \boldsymbol{e}^{t-})$ and $(\boldsymbol{e}^{v-}, \boldsymbol{e}^t)$ are the corresponding hardest negative pairs in a mini-batch.

TABLE I

COMPARISON RESULTS ON THE FLICKR30K DATASET. "*" REPRESENTS THE AVERAGED SIMILARITY SCORES OF TWO SINGLE MODELS TRAINED INDEPENDENTLY. THE BEST AND SECOND-BEST PERFORMANCES ARE IN BOLD AND UNDERLINE, RESPECTIVELY

	Text-to-Image			Image-to-Text			Rsum
Methods	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN* [55]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
CAMP [28]	68.1	89.7	95.2	51.5	77.1	85.3	466.9
VSRN* [7]	71.3	90.6	96.0	54.7	81.8	88.2	482.6
SGM [16]	71.8	91.7	95.5	53.5	79.6	86.5	478.6
MMCA [56]	74.2	92.8	96.4	54.8	81.4	87.8	487.4
VSM [40]	70.8	92.7	96.0	59.5	85.6	91.0	495.6
VSM+Dist [40]	70.8	92.7	96.0	60.9	86.1	91.0	497.5
GSMN* [15]	76.4	94.3	97.3	57.4	82.3	89.0	496.7
CAMERA* [29]	78.0	95.1	**97.9**	60.3	85.9	91.7	508.9
SAF [10]	73.7	93.3	96.3	56.1	81.5	88.0	488.9
NCR [32]	77.3	94.0	97.5	59.6	84.4	89.9	502.7
Ours	78.0	94.6	97.4	59.4	85.8	91.3	506.5
Ours*	**80.2**	**95.3**	**97.9**	**61.2**	**86.8**	**92.0**	**513.4**

Overall, the whole loss function of our method is defined as

$$\mathcal{L} = \lambda \mathcal{L}_g + \mathcal{L}_{WD} + \mathcal{L}_{rank}, \tag{16}$$

where $\mathcal{L}_g$ is the loss function of the improved graph variational autoencoder, $\mathcal{L}_{WD}$ is the Wasserstein distance loss for aligning the graph embedding distributions of images and texts, defined in Eq. 13, $\mathcal{L}_{rank}$ is the ranking loss, and λ is a trade-off parameter.

To learn more discriminative representations, we introduce a two-phase training strategy owing to the fact that the condition of randomly sampled latent distribution is unstable at the beginning of training. In the first phase, the graph construction network is trained by using the ranking loss $\mathcal{L}_{rank}$. In the second phase, the whole network is trained by using the whole loss $\mathcal{L}$. Training the improved graph variational autoencoder is vulnerable to suffer from KL vanishing during training. We use an annealing algorithm to mitigate this issue, and the parameter λ is initialized to 1e-3 and gradually increased to 0.5 until epoch 15 in Flickr30K or epoch 30 in COCO.

IV. EXPERIMENTS

A. Datasets

We evaluate our method on two different datasets: Flickr30K and COCO. The Flickr30K dataset has 31,783 images with five sentences provided for each image. We split the dataset into 29,000 images for training, 1,000 images for validation, and 1,000 images for testing, which is the same splitting as work [23], [24]. The COCO dataset is a very challenging and large-scale dataset, which is split into 113,287 images for training, 5,000 images for validation and the rest 5,000 for evaluation. Each image has five sentences for description.

B. Implementation

We report results by either averaging over 5 folds of 1K test images or directly evaluating on the full 5k test images and use the image embedding module and text embedding module in [29] to extract the initial region features and word features, where the feature dimension is set to 2048. The dimensions of the latent graph representation and the hidden states are 2048. The margin hyper-parameter m in the ranking loss is set to 0.2. The trade-off parameter λ in the whole loss is set to 1. The Adam optimizer [54] is used, the learning rate is set to 0.0001, and the batch size is set to 128.

C. Evaluation Metrics

We evaluate the image-text matching performance by using the proportion of query that matches the correct item in the top-k results, denoted by R@K, K=1, 5, 10. The sum of all R@K is calculated to evaluate the overall matching performance, denoted by Rsum.

D. Comparison with State-of-the-Art

We compare our method with following state-of-the-art methods, SCAN [55], CAMP [28], VSRN [7], SGM [16], MMCA [56], VSM [40], VSM+Dist [40], PVSE [31], GSMN [15], CAMERA [29], SAF [10], and NCR [32]. Table I and Table II show the comparison results on the Flickr30K dataset and the COCO dataset, respectively. Table I and Table II show the comparison results on the Flickr30K dataset and the COCO dataset, respectively, where the integrated model marked "*" represents the average similarity score of the two individual models trained independently, with the best and suboptimal results marked with bold and underlined, respectively. From the results, we have the following observations:

- Our method achieves competitive results in both text-to-image and image-to-text matching on the two datasets, indicating the superiority of learning effective latent graph representation to reduce distractions of entity relationships. The performance improvement on the COCO dataset is not so significant as that on the Flickr30K dataset. The possible reason is that COCO images have fewer objects and simpler relationships, and the advantage of the latent graph representation learning on reducing distractions is not fully exploited.
- Compared with the most relevant methods that also aim to solve the distraction problem (VSM+Dist, PVSE, SAF and NCR), our method generally achieves better

TABLE II

COMPARISON RESULTS ON THE COCO DATASET. "*" REPRESENTS THE AVERAGED SIMILARITY SCORES OF TWO SINGLE MODELS TRAINED INDEPENDENTLY. THE BEST AND SECOND-BEST RESULTS ARE IN BOLD AND UNDERLINE, RESPECTIVELY

	COCO 1K							COCO 5K						
	Text-to-Image			Image-to-Text			Rsum	Text-to-Image			Image-to-Text			Rsum
Methods	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
SCAN* [55]	72.7	94.8	98.4	58.8	88.4	94.8	507.9	50.4	82.2	90.0	38.6	69.3	80.4	410.9
CAMP [28]	72.3	94.8	98.3	58.5	87.9	95.0	506.8	50.1	82.1	89.7	39.0	68.9	80.2	410.0
VSRN* [7]	76.2	94.8	98.2	62.8	89.7	95.1	516.8	53.0	81.1	89.4	40.5	70.6	81.1	415.7
SGM [16]	73.4	93.8	97.8	57.5	87.3	94.3	504.1	50.0	79.3	87.9	35.3	64.9	76.5	393.9
MMCA [56]	74.8	95.6	97.7	61.6	89.8	95.2	514.7	54.0	82.5	90.7	38.7	69.7	80.8	416.4
VSM [40]	77.0	96.1	98.7	65.1	**93.1**	**97.9**	527.9	51.2	81.7	89.1	39.4	72.5	**84.1**	418.0
VSM+Dist [40]	77.8	96.1	98.7	**66.2**	93.0	**97.9**	**529.7**	51.4	81.8	89.1	40.5	**73.5**	**84.1**	420.4
VSM+Dist w/o attribute [40]	74.1	93.2	96.9	61.8	91.5	96.0	513.5	-	-	-	-	-	-	-
PVSE [31]	69.2	91.6	96.6	55.2	86.5	93.7	492.8	45.2	74.3	84.5	32.4	63.0	75.0	374.4
GSMN* [15]	78.4	**96.4**	98.6	63.3	90.1	95.7	522.5	-	-	-	-	-	-	-
CAMERA* [29]	77.5	96.3	98.8	63.4	90.9	95.8	522.7	55.1	82.9	91.2	40.5	71.7	82.5	423.9
SAF [10]	76.1	95.4	98.3	61.8	89.4	95.3	516.3	53.3	-	90.1	39.8	-	80.2	-
NCR [32]	**78.7**	95.8	98.5	63.3	90.4	95.8	522.5	-	-	-	-	-	-	-
Ours	76.1	95.6	98.5	62.4	90.3	95.4	518.3	53.1	82.2	90.5	39.2	70.7	81.5	417.2
Ours*	77.8	96.1	**99.0**	63.9	91.1	96.0	523.9	**55.2**	**83.9**	**91.4**	**40.7**	71.9	82.6	**425.7**

- performance on most evaluation metrics, showing the benefit of the improved graph variational autoencoder on separating the alignable latent factor and the unalignable distracting factors of entity relationships.
- Compared with other methods that introduce complex cross-attention modules (SCAN, CAMP, SGM, MMCA and GSMN), our method only uses a simple inner cosine product to calculate the semantic similarity between image and text, but still achieves better performance, which further validates the importance of disengaging the distracting factor in image-text matching.

E. Ablation Study

To investigate the effectiveness of the components, we introduce several variants of our method for comparison on the Flickr30K and COCO 5K datasets using a single model.

1) Effect of Latent Graph Representation: To verify whether the latent graph representation is profitable for image-text matching, we compare our method with the following variants.

- **Baseline:** we remove the visual and textual graph structures to evaluate the graph representations. In this case, the images and texts are encoded using the embedding modules in [29]. That is, both the graph construction module and the improved graph variational autoencoder are removed.
- **w/o graph:** we replace the improved graph variational autoencoder with a variational autoencoder and remove the graph construction module, to evaluate the effectiveness of latent graph representations.
- **GCN for both image and text:** we use a two-layer GCN to encode both the visual graph and the textual graph, to evaluate the importance of learning edge weights for graph encoding.
- **GNN for both image and text:** we use a two-layer self-attention GNN to encode both the visual and textual graphs. This variant is introduced to evaluate the effectiveness of the combination of a self-attention GNN and a GGNN on encoding the textual graph.
- **GNN+GGNN for both image and text:** we use the combination of a self-attention GNN and a GGNN to encode both the visual and textual graphs. This variant is introduced to evaluate the effectiveness of the two-layer GNN on encoding the visual graph.
- **w/o graph decoders:** we remove both the visual and textual graph decoders to evaluate their effectiveness in learning latent graph representations.
- **w/o prior distribution:** we replace the randomly sampled reparametrized representation with the updated node feature representations, in order to verify the importance of modeling prior distribution in learning latent graph representations.
- **w/o cross-attention:** we remove the cross-attention module to evaluate its effectiveness in learning latent graph representations.

The results are shown in Table III and we make the following observations:

- "w/o graph" generally performs worst when compared with other variants, since it dose not exploit the semantic relationships between entities, which shows that modeling entity relationships facilitates matching betwee images and texts.
- The performance of "GCN for both image and text" dramatically degrades compared with our method, which clearly shows that it is important to learn the weights of different neighborhoods by using self-attention GNN in graph encoding.
- Our method achieves better results than "GNN for both image and text", showing that it is beneficial to encode the textual graph by combining of a self-attention GNN and a GGNN. The reason is that texts have the property of sequence and GGNN is more suitable for updating previous states in early iterations when encoding the text data.

TABLE III

ABLATION ANALYSIS OF DIFFERENT COMPONENTS ON THE FLICKR30K AND COCO 5K DATASETS

	Flickr30K						COCO 5K					
	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
Methods	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	75.8	93.4	96.7	58.1	84.5	90.6	49.6	79.2	88.5	36.5	67.8	79.2
w/o graph	73.3	93.3	96.6	55.8	82.9	89.5	45.8	76.1	86.8	33.9	64.5	76.3
GCN for both image and text	74.4	94.0	96.9	57.0	83.3	90.3	48.2	78.8	88.2	36.4	67.5	79.1
GNN for both image and text	77.0	94.5	97.1	58.0	84.7	90.3	52.8	81.9	90.0	38.8	69.9	81.0
GNN+GGNN for both image and text	75.6	94.2	97.0	58.4	84.9	91.0	45.5	76.0	86.1	34.8	66.0	77.7
w/o graph decoders	77.0	94.3	97.1	58.5	84.8	91.0	53.1	81.5	90.0	38.7	69.9	81.2
w/o prior distribution	76.3	94.2	97.1	58.9	85.3	90.9	52.6	81.9	90.4	38.9	69.9	81.0
w/o cross-attention	77.8	**94.6**	97.2	58.2	85.2	90.6	**53.2**	81.8	89.9	38.7	69.8	80.7
Ours	**78.0**	**94.6**	**97.4**	**59.4**	**85.8**	**91.3**	53.1	**82.2**	**90.5**	**39.2**	**70.7**	**81.5**

- Our method also achieves better results than "GNN+GGNN for both image and text", showing that the two-layer GNN is more suitable for encoding the image data.
- When removing the textual graph decoder and the visual graph encoder in "w/o graph decoders", the performance drops, which indicates that the graph decoder acts as a sieve to capture necessary information of input data, encouraging the approximate posterior estimation more accurate.
- "w/o prior distribution" achieves a slight drop in the matching performance, probably due to the lack of regularity of the learned latent space, leading to the decoding of meaningless data.
- When removing the cross-attention module in "w/o cross-attention", the performance has a significant drop, which shows the importance of building the interactions between different modalities to promote learning the cross-modality consistency.

2) Effect of Loss Function: To evaluate the impact of each loss function, we compare our method with the following variants.

- **w/o $\mathcal{L}_g$:** we remove the loss function of the improved graph variational autoencoder $\mathcal{L}_g$ to evaluate its effectiveness, and we only use $\mathcal{L}_{WD}$ and $\mathcal{L}_{rank}$ for training.
- **w/o $\mathcal{L}_{WD}$:** we remove the Wasserstein distance loss $\mathcal{L}_{WD}$ to evaluate whether aligning the distributions of latent semantic representations across modalities is useful to the quality of disentangled representations.
- **replace $\mathcal{L}_{WD}$ with $\mathcal{L}_{KD}$:** we replace $\mathcal{L}_{WD}$ with $\mathcal{L}_{KD}$ to evaluate the effectiveness of minimizing KL divergence between the distributions of latent representations for both modalities on learning latent graph representations.

Table IV reports the results on the Flickr30K and COCO 5K datasets. For "w/o $\mathcal{L}_g$", it is obvious that without the supervision of the improved graph variational autoencoder loss function, our method fails to disentangle the representations in each modality. For "w/o $\mathcal{L}_{WD}$", we observe that removing Wasserstein distance loss results in relatively low performance, which validates that the distance loss has a positive impact on training the network and guides the model to make the projections of data points from the visual and textual modalities to be closer in the learned latent space. For "replace $\mathcal{L}_{WD}$ with $\mathcal{L}_{KD}$", we observe that the performance usually degrades, due to that our method easily suffers from KL-vanishing problem by computing KL divergence between the distributions of latent representations across modalities.

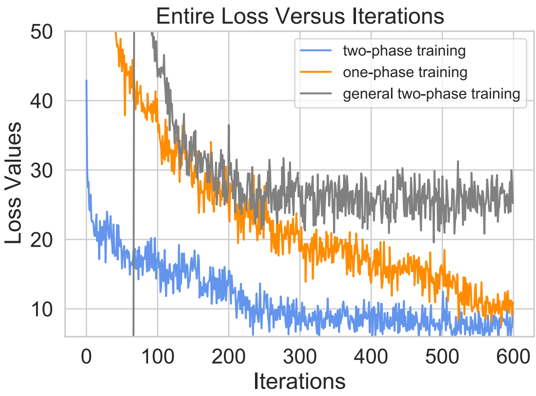


Fig. 3. Comparison of loss curves with different training strategies on the Flickr30K dataset.

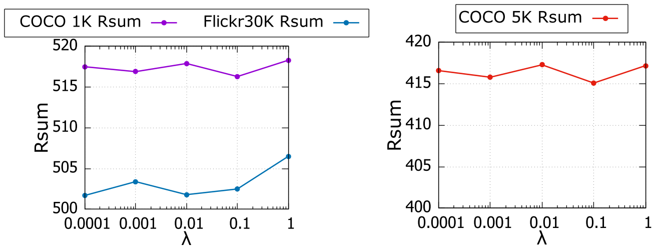


Fig. 4. Comparison results of different trade-off parameters on the Flickr30K and COCO datasets.

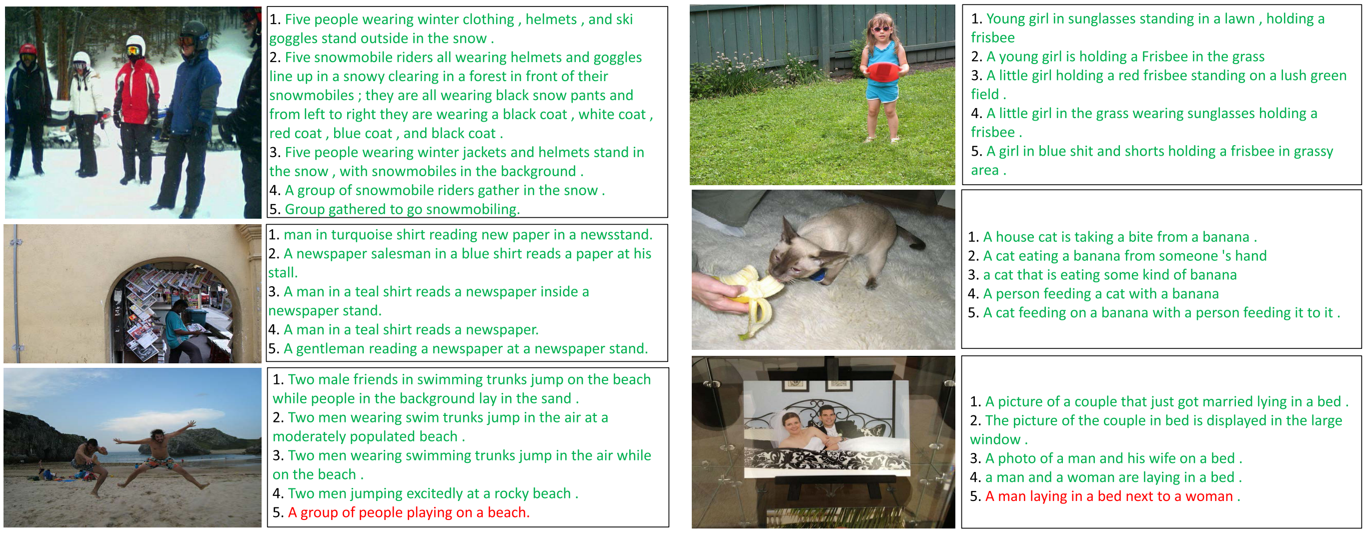
3) Effect of Training Strategy: To evaluate the effectiveness of the proposed two-phase training strategy, we compare our method with following variants.

- **general two-phase training:** we first pre-train the network using the embedding loss $\lambda\mathcal{L}_g + \mathcal{L}_{WD}$, and then fine-tune it using the whole loss $\mathcal{L}$.
- **one-phase training:** we train the network in one phase.

Table IV reports the results on the Flickr30K and COCO 5K datasets. For "general two-phase training strategy", we observe that the matching performance degrades significantly. The

TABLE IV
ABLATION ANALYSIS OF DIFFERENT LOSSES AND TRAINING STRATEGIES ON THE FLICKR30K AND COCO 5K DATASETS

	Flickr30K						COCO 5K					
	Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text		
Methods	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
w/o $\mathcal{L}_g$	77.4	93.7	97.2	58.7	85.0	90.8	52.8	81.4	90.4	38.7	70.0	81.4
w/o $\mathcal{L}_{WD}$	77.3	94.7	97.0	58.5	84.9	90.7	50.6	79.4	88.9	36.9	68.2	79.8
replace $\mathcal{L}_{KD}$ with $\mathcal{L}_{WD}$	76.7	**94.9**	96.9	58.1	84.6	91.0	52.2	81.7	89.8	39.0	70.2	81.3
general two-phase training	56.8	83.6	91.6	43.8	74.7	83.5	47.7	78.1	87.5	35.6	67.1	78.7
one-phase training	77.2	94.0	97.0	58.5	84.3	90.9	52.9	81.4	90.2	39.1	70.2	81.4
Ours	**78.0**	94.6	**97.4**	**59.4**	**85.8**	**91.3**	**53.1**	**82.2**	**90.5**	**39.2**	**70.7**	**81.5**



1. Five people wearing winter clothing , helmets , and ski goggles stand outside in the snow .
2. Five snowmobile riders all wearing helmets and goggles line up in a snowy clearing in a forest in front of their snowmobiles ; they are all wearing black snow pants and from left to right they are wearing a black coat , white coat , red coat , blue coat , and black coat .
3. Five people wearing winter jackets and helmets stand in the snow , with snowmobiles in the background .
4. A group of snowmobile riders gather in the snow .
5. Group gathered to go snowmobiling.

1. man in turquoise shirt reading new paper in a newsstand.
2. A newspaper salesman in a blue shirt reads a paper at his stall.
3. A man in a teal shirt reads a newspaper inside a newspaper stand.
4. A man in a teal shirt reads a newspaper.
5. A gentleman reading a newspaper at a newspaper stand.

1. Two male friends in swimming trunks jump on the beach while people in the background lay in the sand .
2. Two men wearing swim trunks jump in the air at a moderately populated beach .
3. Two men wearing swimming trunks jump in the air while on the beach .
4. Two men jumping excitedly at a rocky beach .
5. A group of people playing on a beach.

(a) Flickr30K

1. Young girl in sunglasses standing in a lawn , holding a frisbee
2. A young girl is holding a Frisbee in the grass
3. A little girl holding a red frisbee standing on a lush green field .
4. A little girl in the grass wearing sunglasses holding a frisbee .
5. A girl in blue shit and shorts holding a frisbee in grassy area .

1. A house cat is taking a bite from a banana .
2. A cat eating a banana from someone 's hand
3. a cat that is eating some kind of banana
4. A person feeding a cat with a banana
5. A cat feeding on a banana with a person feeding it to it .

1. A picture of a couple that just got married lying in a bed .
2. The picture of the couple in bed is displayed in the large window .
3. A photo of a man and his wife on a bed .
4. a man and a woman are laying in a bed .
5. A man laying in a bed next to a woman .

(b) COCO

Fig. 5. Top-5 qualitative results of image-to-text retrieval on the Flickr30K and COCO datasets. The ground-truth and mismatched sentences are in green and red, respectively.

possible reason is that the condition of randomly sampled latent distribution is unstable at the beginning of training, leading to a failure in learning a good common latent space. For "one-phase training strategy", it is obvious that our two-phase training strategy significantly improves the matching performance. Moreover, Fig. 3 demonstrates that the our training strategy has a more apparent downward trend and reaches the convergence state within 15 epoch.

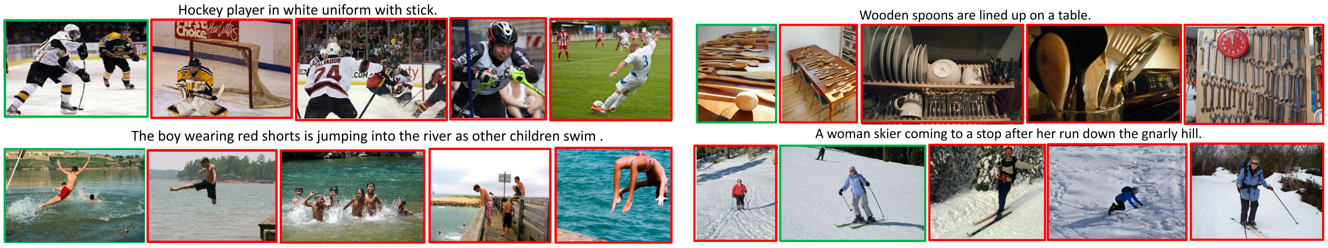
4) Effect of the Hyperparameter λ*:* We evaluate the impact of the trade-off hyperparameter λ in Eq. 16 by tuning its values from $\{0.0001, 0.001, 0.01, 0.1, 1\}$. The results are shown in Fig. 4. It is obvious that our method achieves the best performance when $\lambda = 1$ on the two datasets.

F. Qualitative Results

1) Visualization of Image-Text Matching Results: To further demonstrate the effectiveness of our method, we visualize several examples of image-to-text retrieval results on the two datasets in Fig. 5, where the correctly matched sentences are shown in green and the unmatched ones are shown in red. We also visualize several examples of text-to-image retrieval in Fig. 6, where the true matches are shown in green boxes while the mismatches are in red. Note that in our settings, we rank the top-5 retrieval results according to the similarity scores between images and sentences, where each image has five corresponding sentences and each sentence has only one corresponding image.

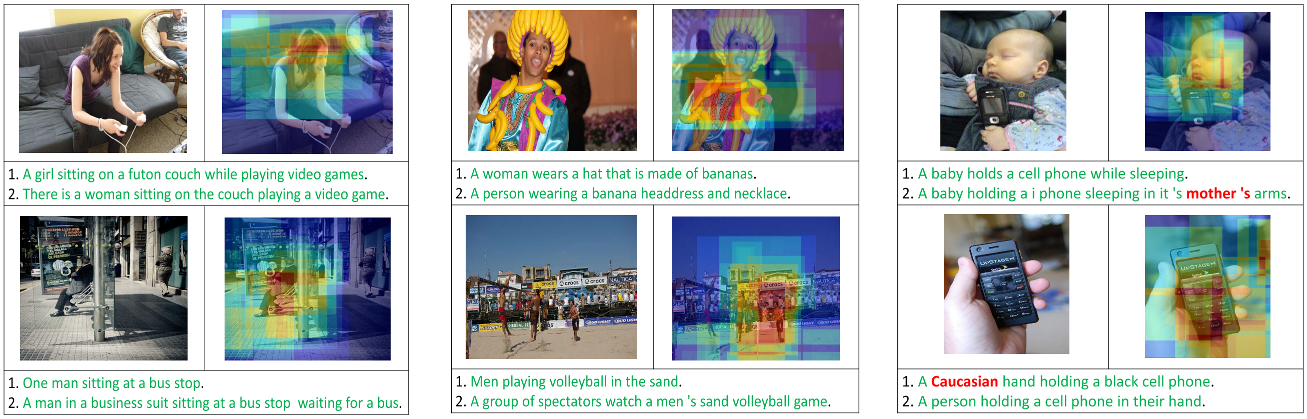
From Fig. 5, we observe that almost all the paired sentences have been successfully retrieved by using our method, except for the retrieved sentences in red at the bottom line that do not correspond to the image, but are still semantically related to the ground-truth sentence. As can be seen from Fig. 6, our method completes accurate matching for almost all correct sentences, owing to its ability of effectively reducing distractions and decreasing the ranking of unmatched samples. For the second example in Fig. 6 (b), we note that the query sentence "A woman skier coming to a stop after her run down the gnarly hill" whose corresponding top-1 retrieved image is wrong, but its semantics are almost consistent with the correct image in green.

2) Visualization of Distracting Factors: To gain deep insights into the impact of distracting factors of entity relationships, we visualize several typical image-to-text retrieval examples in Fig. 7. From these results, it is interesting to observe that: (1) as shown in Fig. 7 (a), the "woman sitting"


Hockey player in white uniform with stick.
The boy wearing red shorts is jumping into the river as other children swim .
Wooden spoons are lined up on a table.
A woman skier coming to a stop after her run down the gnarly hill.

(a) Flickr30K (b) COCO

Fig. 6. Top-5 qualitative results of text-to-image retrieval on the Flickr30K and COCO datasets. The ground-truth and mismatched images are shown in green and red boxes, respectively.



1. A girl sitting on a futon couch while playing video games.
2. There is a woman sitting on the couch playing a video game.

1. One man sitting at a bus stop.
2. A man in a business suit sitting at a bus stop waiting for a bus.

1. A woman wears a hat that is made of bananas.
2. A person wearing a banana headdress and necklace.

1. Men playing volleyball in the sand.
2. A group of spectators watch a men 's sand volleyball game.

1. A baby holds a cell phone while sleeping.
2. A baby holding a i phone sleeping in it 's mother 's arms.

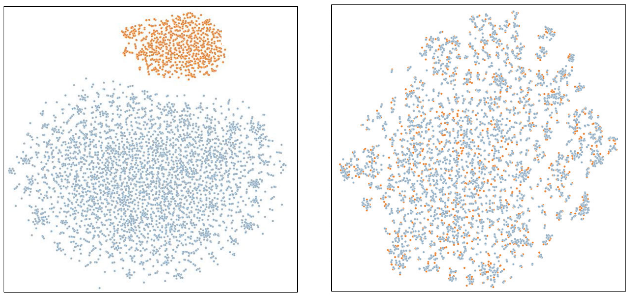
1. A Caucasian hand holding a black cell phone.
2. A person holding a cell phone in their hand.

(a) Irrelevant object regions (b) Irrelevant background regions (c) Regions corresponding to noisy words

Fig. 7. Visualization of attention weights of each image region on image-text matching.

receives more visual attention than the other person, and the "man sitting" attracts more than the pedestrian. Both the other person and the pedestrian are irrelevant object regions and thus create distractions on image-text matching. This demonstrates that our method succeeds in reducing the distractions from irrelevant object regions; (2) as shown in Fig. 7 (b), the salient foreground regions get attention rather than rather than the trivial background regions, verifying the advantage of our method on reducing the distractions from irrelevant background regions; (3) as shown in Fig. 7 (c), the words "mother" and "caucasian" are invisible in images and obtain less visual attention, demonstrating the ability of our method on reducing the distractions from noisy words.

Overall, these retrieval examples suggest that our method learns robust latent graph representations by reducing the distractions of irrelevant or misleading information for image-text matching.



(a) Modality-specific space (b) Learned latent space

Fig. 8. T-SNE visualized results of (a) modality-specific space and (b) latent graph space on the Flickr30K test set. The orange points and gray points denote the image and text, respectively.

3) Visualization of Latent Graph Representation Space: To demonstrate the learned latent graph representation space, we depict the T-SNE visualized results on the Flickr30K dataset in Fig. 8. Specifically, Fig. 8 (a) shows the modality-specific space, where the private representations (the representations after taking readout operation on $\boldsymbol{\sigma}^v$ and $\boldsymbol{\sigma}^t$) within each modality are learned. Fig. 8 (b) shows the latent space where the latent representations ($\boldsymbol{e}^v$ and $\boldsymbol{e}^t$) across modalities are learned. It is interesting to observe that in the latent representation space that the paired image and text are closer to each other, demonstrating that the distracting factors are effectively excluded to narrow the heterogeneity gap between image and text.

V. Conclusion

We have presented an adaptive latent graph representation learning method for image-text matching. An improved graph variational autoencoder is used to disentangle distracting factors for learning latent graph representations of different modalities, achieving robust cross-modal matching. An adaptive latent graph attention module is introduced to focus more on salient and common representations of relationships, succeeding in narrowing the heterogeneous gap between different modalities. We use a two-phase training strategy that can improve the discrimination of multi-modal feature representation, thus further boosting the image-text matching performance. In the future, we are going to extend the latent graph representation learning method to the video-text matching task.

References

[1] S. Chun, S. J. Oh, R. S. de Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8415–8424.

[2] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020.

[3] W. Zhao, X. Wu, and J. Luo, "Cross-domain image captioning via cross-modal retrieval and model adaptation," *IEEE Trans. Image Process.*, vol. 30, pp. 1180–1192, 2021.

[4] S. Wang, Z. Yao, R. Wang, Z. Wu, and X. Chen, "FAIEr: Fidelity and adequacy ensured image caption evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14050–14059.

[5] L. V. B. Beltrán, J. C. Caicedo, N. Journet, M. Coustaty, F. Lecellier, and A. Doucet, "Deep multimodal learning for cross-modal retrieval: One model for all tasks," *Pattern Recognit. Lett.*, vol. 146, pp. 38–45, Jun. 2021.

[6] M. Lao, Y. Guo, N. Pu, W. Chen, Y. Liu, and M. S. Lew, "Multi-stage hybrid embedding fusion network for visual question answering," *Neurocomputing*, vol. 423, pp. 541–550, Jan. 2021.

[7] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4654–4662.

[8] J. Wehrmann, C. Kolling, and R. C. Barros, "Adaptive cross-modal embeddings for image-text alignment," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 7, 2020, pp. 12313–12320.

[9] Z. Ji, H. Wang, J. Han, and Y. Pang, "Saliency-guided attention network for image-sentence matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5754–5763.

[10] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 1218–1226.

[11] E. Mavroudi, B. B. Haro, and R. Vidal, "Representation learning on visual-symbolic graphs for video understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Mar. 2020, pp. 71–90.

[12] W. Zhang et al., "Relational graph learning for grounded video description generation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3807–3828.

[13] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7370–7377.

[14] X. Liu, X. You, X. Zhang, J. Wu, and P. Lv, "Tensor graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 8409–8416.

[15] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10921–10930.

[16] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1508–1517.

[17] S. Long, S. C. Han, X. Wan, and J. Poon, "Gradual: Graph-based dual-modal representation for image-text matching," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 3459–3468.

[18] S.-J. Peng, Y. He, X. Liu, Y.-M. Cheung, X. Xu, and Z. Cui, "Relation-aggregated cross-graph correlation learning for fine-grained image–text retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 13, 2022, doi: 10.1109/TNNLS.2022.3188569.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–14.

[20] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4107–4116.

[21] A. Frome et al., "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 2121–2129.

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.

[23] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.

[24] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[25] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2310–2318.

[26] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6077–6086.

[27] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 201–216.

[28] Z. Wang et al., "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5764–5773.

[29] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1–9.

[30] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, "Dynamic modality interaction modeling for image-text retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2021, pp. 1104–1113.

[31] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1979–1988.

[32] Z. Huang et al., "Learning with noisy correspondence for cross-modal matching," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[33] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10685–10694.

[34] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 211–229.

[35] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2615–2624.

[36] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 11021–11028.

[37] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, "Language-conditioned graph networks for relational reasoning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10294–10303.

[38] W. Yu, J. Zhou, W. Yu, X. Liang, and N. Xiao, "Heterogeneous graph learning for visual commonsense reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 2769–2779.

[39] A. Zareian, Z. Wang, H. You, and S.-F. Chang, "Learning visual commonsense for robust scene graph generation," in *Computer Vision*. Glasgow, U.K.: Springer, Aug. 2020, pp. 642–657.

[40] Y. Li, D. Zhang, and Y. Mu, "Visual-semantic matching by exploring high-order attention and distraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12786–12795.

[41] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero- and few-shot learning via aligned variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8247–8255.

[42] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 4847–4857.

[43] J. Li, M. Jing, L. Zhu, Z. Ding, K. Lu, and Y. Yang, "Learning modality-invariant latent representations for generalized zero-shot learning," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1348–1356.

[44] M. Ye and J. Shen, "Probabilistic structural latent representation for unsupervised embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5457–5466.

[45] J. Wei, Y. Yang, X. Xu, Y. Ji, X. Zhu, and H. T. Shen, "Graph-based variational auto-encoder for generalized zero-shot learning," in *Proc. 2nd ACM Int. Conf. Multimedia Asia*, 2021, pp. 1–7.

[46] R. Keshari, R. Singh, and M. Vatsa, "Generalized zero-shot learning via over-complete distribution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13300–13308.

[47] X. Wen, Z. Han, and Y.-S. Liu, "CMPD: Using cross memory network with pair discrimination for image-text retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2427–2437, Jun. 2021.

[48] M. Jing, J. Li, L. Zhu, K. Lu, Y. Yang, and Z. Huang, "Incomplete cross-modal retrieval with dual-aligned variational autoencoders," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3283–3291.

[49] H. Fu, R. Wu, C. Liu, and J. Sun, "MCEN: Bridging cross-modal gap between cooking recipes and dish images with latent variable model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14570–14580.

[50] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Trans. Multimedia*, vol. 24, pp. 1763–1774, 2021.

[51] M. Kim, R. Guerrero, and V. Pavlovic, "Learning disentangled factors from paired data in cross-modal retrieval: An implicit identifiable VAE approach," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2862–2870.

[52] W.-N. Hsu and J. Glass, "Disentangling by partitioning: A representation learning framework for multimodal sensory data," 2018, *arXiv:1805.11264*.

[53] Y. Fujiwara et al., "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

[54] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.

[55] R. Zhao, K. Zheng, and Z.-J. Zha, "Stacked convolutional deep encoding network for video-text retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[56] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10941–10950.

Mengxiao Tian (Student Member, IEEE) received the B.S. degree from Southwest Minzu University, Chengdu, China, in 2017, and the M.S. degree from China Agricultural University, Beijing, China, in 2020. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include computer vision, machine learning, and vision-language.

Xinxiao Wu (Member, IEEE) received the B.S. degree in computer science from the Nanjing University of Information Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in computer science from the Beijing Institute of Technology (BIT), Beijing, China, in 2010. From 2010 to 2011, she was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. She is currently a Full Professor with the School of Computer Science, BIT. Her research interests include computer vision, machine learning, and vision-language. She serves on the Editorial Board for the IEEE TRANSACTIONS ON MULTIMEDIA.

Yunde Jia (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechatronic engineering from the Beijing Institute of Technology (BIT). He was a Visiting Scientist with The Robotics Institute, Carnegie Mellon University (CMU), from 1995 to 1997. He was a Professor of computer science at BIT from 2000 to 2022, and directed the Beijing Key Laboratory of Intelligent Information Technology from 2001 to 2022. He is currently a Chair Professor of computer science at Shenzhen MSU-BIT University, and serves as the Director of the Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing. His interests include computer vision, computational perception and cognition, and intelligent systems.