

# Cross-modal Image Retrieval with Deep Mutual Information Maximization

Chunbin Gu<sup>a,c</sup>, Jiajun Bu<sup>a,c,d,\*</sup>, Xixi Zhou<sup>a,c</sup>, Chengwei Yao<sup>a,c</sup>, Dongfang Ma<sup>b,e</sup>, Zhi Yu<sup>a,c</sup>, Xifeng Yan<sup>f</sup>

<sup>a</sup>*Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, 310007, Hangzhou, P.R.China*

<sup>b</sup>*Institute of Marine Sensing and Networking, Zhejiang University, 310058, Hangzhou, P.R.China*

<sup>c</sup>*Alibaba-Zhejiang University Joint Institute of Frontier Technologies, 310007, Hangzhou, P.R.China*

<sup>d</sup>*MOE Key Laboratory of Machine Perception, 100871, Beijing, P.R.China*

<sup>e</sup>*Key Laboratory of Ocean Observation-Imaging Testbed of Zhejiang Province, Zhejiang University, Zhoushan, 316021, P.R.China*

<sup>f</sup>*Department of Computer Science, University of California, Santa Barbara, CA, 93106*

---

## Abstract

In this paper, we study the cross-modal image retrieval, where the inputs contain a source image plus some text that describes certain modifications to this image and the desired image. Prior work usually uses a three-stage strategy to tackle this task: 1) extract the features of the inputs; 2) fuse the feature of the source image and its modified text to obtain fusion feature; 3) learn a similarity metric between the desired image and the source image + modified text by using deep metric learning. Since classical image/text encoders can learn the useful representation and common pair-based loss functions of distance metric learning are enough for cross-modal retrieval, people usually improve retrieval accuracy by designing new fusion networks. However, these methods do not successfully handle the modality gap caused by the inconsistent distribution and representation of the features of different modalities, which greatly influences the feature fusion and the similarity learning. To alleviate this problem,

---

\*Corresponding author

*Email addresses:* guchunbin@zju.edu.cn (Chunbin Gu), bjj@zju.edu.cn (Jiajun Bu), xixi.zxx@zju.edu.cn (Xixi Zhou), yaochw@zju.edu.cn (Chengwei Yao), mdf2004@zju.edu.cn (Dongfang Ma), yuzhirenzhe@zju.edu.cn (Zhi Yu), xyan@cs.ucsb.edu (Xifeng Yan)

we adopt the contrastive self-supervised learning method Deep InforMax (DIM) [1] to our approach to bridge this gap by enhancing the dependence between the text, the image, and their fusion. Specifically, our method narrows the modality gap between the text modality and the image modality by maximizing mutual information between their not exactly semantically identical representation. Moreover, we seek an effective common subspace for the semantically same fusion feature and desired image’s feature by utilizing Deep InforMax between the low-level layer of the image encoder and the high-level layer of the fusion network. Extensive experiments on three large-scale benchmark datasets show that we have bridged the modality gap between different modalities and achieve state-of-the-art retrieval performance.

*Keywords:* Cross-modal Image Retrieval, Mutual Information, Deep Metric Learning, Self-supervised Learning

---

## 1. Introduction

Image retrieval is a key compute vision problem and it has made great progress due to deep learning [2, 3, 4, 5, 6]. Cross-modal image retrieval allows using other types of query, such as text to image retrieval [7, 8], sketch to image retrieval [9, 10] and cross-view image retrieval [11, 12]. In this paper, we consider the case where input queries are formulated as an input image plus the text that describes desired modifications to the image. Different from attribute-based image retrieval [13], our input text can be multi-word instead of a single attribute. For instance, our input image is a women clog and the text could be “have a buckle and strap, no patterns”. Our desired image should meet the requirement of the two input modalities (Figure 1).

To solve this problem, TIRG [14] first extracts the features of the source image and the modification text by ResNet-18 and LSTM respectively, then fuses them via a gated residual connection, and finally learns the similarity metric using softmax cross-entropy loss to minimize the distance between fusion feature and the feature of the desired image. Similarly, another research work

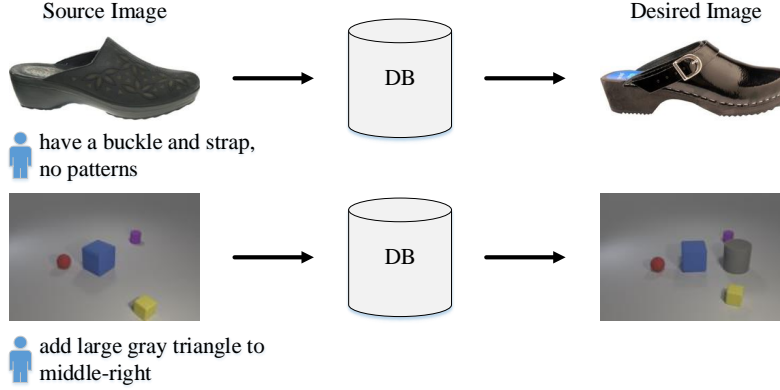


Figure 1: Example of image retrieval based on image and text fusion. The text states the desired modification to the image and the information of the two input modalities conveys into the system.

[15] gets the feature of the source image and that of the text by ResNet-101 and GRU, fuses them through concatenation and improves the rank of the desired image by deep metric learning based on reinforcement learning. Although there are many other related work [16, 17, 18, 19], most of them focus on designing new feature fusion networks and employing different DML loss function. These methods align the feature vector and assess the similarity between the desired image and the fusion of the source image and the modified text by common retrieval losses. However, since features of different modalities usually have inconsistent distribution and representation, a modality gap exists so as to affect the retrieval performance significantly [20].

Mutual information (MI) can capture non-linear statistical dependencies between random variables and act as a measure of true dependence [21]. The recent research [22, 23, 1] offers various general-purpose parametric neural estimators of mutual information between different representations in the deep neural network. Thus, we align the feature distributions of text, image, and their fusion by Deep InfoMax (DIM) [1] between the representations in the encoders of these modalities. Specifically, we maximize the MI between the low-level representation in the text encoder and the high-level representation in the desired image

encoder (ITDIM) to project these two modalities into a common subspace. As the text and the desired image are not exactly semantically same, we realize ITDIM by estimating their overlapping semantic information. Compared with two modalities with independent distribution (like text and images), the image modality is the key component of the fusion modality, so their features' distribution aligns with each other to some extent. Our method gets a better alignment by maximizing the MI between the low-level representation of the desired image and the fusion's high-level representation (IFDIM). Here the semantic information of the different-level representations is identical. A handful of literature in the text to image retrieval field narrows the modality gap using adversarial loss [20, 24], which attempts to make the features of different modalities indistinguishable. In essence, these methods can be treated as the special cases of MINE (the basic version of Deep InfoMax) [22], which maximize the MI between the last layers of different encoders using minimax objective [22, 1, 25]. Maximizing mutual information between two complete feature vectors is often insufficient for increasing the dependence of two modalities. Therefore, DIM also maximizes the average MI between the high-level representation and local regions of the low-level representation (e.g., patches rather than the complete image) to make the alignment better. Because our method uses the Text Image Residual Gating (TIRG) [14] as our basic network architecture, we call it TIRG-DIM.

The experiment shows that the proposed method can achieve higher retrieval accuracy compared to existing methods on three standard benchmark datasets, namely Fashion200K [26], MIT-states [27] and CSS [14].

To summarize, our contributions are threefold:

- We design a novel framework for cross-modal image retrieval based on Deep InfoMax. By using ITDIM, maximizing MI by estimating the overlapping semantic information between the representations of the text modality and the image modality, we project the features of these two semantically different modalities with independent distribution into a common

subspace, which can improve retrieval accuracy by learning a higher quality fusion feature.

- We accurately align the distribution of the features of the fusion modality and its main component, the image modality, by IFDIM that maximizes mutual information between the semantically same representations in the fusion network and the desired image encoder, which leads to more competitive retrieval results.
- The empirical results show that our method outperforms the state-of-the-art approaches for cross-modal image retrieval on three public benchmarks, Fashion200K, MIT-states and CSS.

## 2. Related Work

In this section, we briefly review the methods of cross-modal image retrieval based on feature fusion and concisely introduce deep mutual information maximization.

### 2.1. Cross-modal Image Retrieval Based on Feature Fusion

In addition to the methods mentioned before, the previous work [14] also provides seven benchmarks which use the same system pipeline as TIRG except feature fusion modules. For similarity, we define the feature of the source image, that of the modified text and the fusion feature as  $\phi_s$ ,  $\phi_t$  and  $\phi_{st}$ . The feature fusion methods of these benchmarks are as follows,

- Image Only:  $\phi_{st} = \phi_s$ .
- Text Only:  $\phi_{st} = \phi_t$ .
- Concatenating features of image and text using  $\phi_{st} = f_{MLP}([\phi_s, \phi_t])$  [16, 28]. In experiments, it is implemented by making use of two layers of MLP with RELU, the batch-norm and the dropout rate of 0.1.

- Show and Tell [29]: In this method,  $\phi_{st}$  is the final state of a LSTM which encoders the image and the words in the text in turn.
- Attribute as operator: Embed each text as a transformation and apply it to  $\phi_s$  to obtain  $\phi_{st}$  [17]
- Parameter Hashing [18]:  $\phi_{st}$  is the output of the image CNN which replaces the weights of a fc layer with transformation matrix, i.e. the hash of  $\phi_t$ .
- Relationship [30] first constructs relationship features by concatenating text feature  $\phi_t$  and the feature-map vectors from the convolved image; then these features pass through a MLP and the result is averaged to get  $\phi_{st}$ .
- FiLM [19] outputs  $\phi_{st}$  by a feature-wise affine transformation of the image feature,  $\phi_{st} = \gamma_i \phi_s + \beta_i$ , where  $\gamma_i, \beta_i \in R^C$  is the modulation features predicted by  $\phi_t$ , the  $i$  is the index of the layer and  $C$  is the number of features or feature maps.

The above approaches learn the text feature, the desired image feature, and the fusion feature separately. This leads to the modality gap due to the inconsistent distribution of these features, which greatly affects the retrieval accuracy. To alleviate this problem, we align these distributions by maximizing the mutual information between the representations of different modalities.

## 2.2. Deep Mutual Information Maximization

Mutual information (MI) is a fundamental quantity across data science for measuring the relationship between random variables [31, 32, 33]. Unlike correlation, MI captures non-linear statistical dependence between variables and thus can act as a measure of true dependence [21]. Though the infomax principle that the idea of maximizing MI between the input and output has been proposed in many traditional feature learning methods [34, 35], MI is often hard

to compute [36], especially for the high-dimensional and continuous variables in the deep neural network.

Fortunately, the recent research makes a theoretical breakthrough in deep mutual information estimation and provides the method for computing and optimizing the MI between input and output in a deep neural network. Mutual Information Neural Estimation (MINE) [22] is the first general-purpose estimator of the MI of continuous variables. Furthermore, Deep InfoMax [1] leverages local structure apart from global MI utilized in MINE to improve the suitability of representations for classification and provides various MI estimators. Moreover, mutual information maximization between features extracted from multiple views also draws much attention [23, 37], and these studies demonstrate that the quality of the representation improves as the number of views increases. As a member of self-supervised learning [38, 39, 40, 41], deep mutual information maximization exploiting dual optimization to estimate divergences goes beyond the minimax objective as formalized in GANs [42, 43, 44]. Many deep learning tasks have adopted this method for estimating MI via back-propagation and proven its effectiveness, like text generation [45, 46, 47] and representation learning [48, 49, 50].

In the cross-modal field, the use of deep mutual information is diverse. Since the mutual information between different modalities usually has higher semantic meaning compared to information that is modality-specific, people verify whether two input data correspond to each other by capturing mutual information between the two modalities [51, 52]. Also, the researchers utilize mutual information estimation to improve the qualities of representations [53, 54]. In the visual question answering field, Information Maximization Visual Question Generator [55] employs mutual information maximization to guarantee relevance between the generated question with the image and the expected answer. To our best knowledge, there are few research utilizing deep mutual information maximization to align the features' distributions from different modalities up till now.

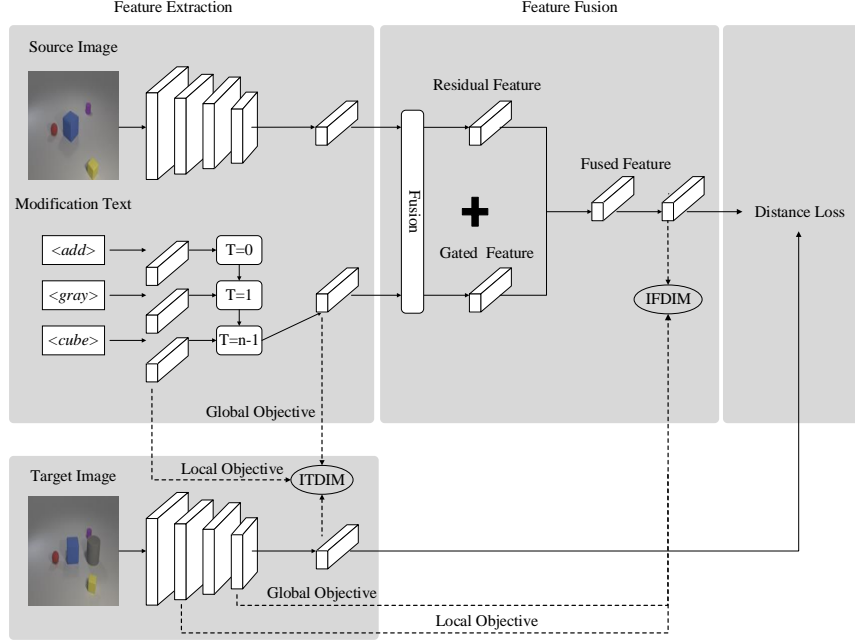


Figure 2: The system pipeline for training based on the abstract modification text.

### 3. Text Image Residual Gating Based on Deep Information Maximization

Since the modality gap caused by the distributional difference of features between modalities significantly influences the cross-modal image retrieval accuracy, we erase this gap by applying mutual information maximization between the representations of the image, the text and their fusion. Figure 2 is the system pipeline of our model.

#### 3.1. Feature Fusion Based on Deep Mutual Information Maximization

In our task, there are two main modality gaps influencing our retrieval accuracy: 1) the gap between the source image and text which makes feature fusion insufficient; 2) the gap between the fusion and the desired image which directly affects the similarity learning. We utilize deep mutual information to narrow these two gaps in this and the next section.



The main task of the feature fusion module is to compose the semantic information extracted from the source image and the modified text. Since the fusion network’s inputs are the features of the source image and the modified text, we first encode each input by a corresponding classical encoder.

For the input images, we adopt a ResNet-18 whose output dimension of the last fc is changed to 512 to extract their features. For the modified text, we firstly embed it into a distributed embedding space and get the word embeddings; Then, we employ a widely used sequence learning model: Long Short-Term Memory (LSTM) to learn the sentence representations. We define the text’s feature as the hidden state at the final time step.

After obtaining these features, we fuse them by the basic network of our method, Text Image Residual Gating (TIRG) [14]. TIRG is composed of gating feature and residual feature and it fuses image and text features by the following approach,

$$\phi_{st}^{rg} = w_g f_g(\phi_s, \phi_t) + w_r f_r(\phi_s, \phi_t) \quad (1)$$

where  $f_g, f_r$  denote gating and the residual features presented in Figure 2.  $w_g$  and  $w_r$  are the trade-off between these two features. This gating feature in TIRG can be formulated as follows,

$$f_g(\phi_s, \phi_t) = \sigma(\text{FC}_{g2}(\text{ReLU}(\text{FC}_{g1}([\phi_s, \phi_t]))) \odot \phi_s \quad (2)$$

where  $\sigma$  is the sigmoid function,  $\odot$  is element wise product,  $\text{FC}_{g1}$  and  $\text{FC}_{g2}$  represent fully connected layers and  $[\phi_s, \phi_t]$  denotes the concatenation of  $\phi_s$  and  $\phi_t$ . This feature is utilized to judge whether the modified text is helpful to the query image or not and retain the image feature if these two inputs are sufficiently different. The residual feature is computed as follows,

$$f_r(\phi_s, \phi_t) = \text{FC}_{r2}(\text{ReLU}(\text{FC}_{r1}([\phi_s, \phi_t]))) \quad (3)$$

where  $\text{FC}_{r1}$  and  $\text{FC}_{r2}$  are fully connected layers.

Figure 2 presents the model for the dataset with abstract modification text, such as Fashion200k and MIT-states, and we call it  $\text{TIRG}_A$ . For the dataset

whose modification text is more concrete like CSS, we use the  $\text{TIRG}_C$  model.  $\text{TIRG}_C$  changes the feature extraction module to alter the spatial properties of the output of the image encoder. It replaces the source image encoder with ResNet-17 and broadcasts the text’s feature along the height and width dimension to make it match the source image’s feature. Accordingly, fully connected layers in the fusion network are replaced by convolutional layers with  $3 \times 3$  filter kernels.

In our framework, distance metric learning optimizes the whole network by measuring the similarity between the desired image and the fusion of the source image and the modified text. As a classical encoder, ResNet-18 can get the desired image’s high-quality feature in the fully labeled dataset [56, 57]. Thus, the quality of the fusion feature is crucial to the retrieval accuracy. However, as the source image and the modified text are from different modalities, their features usually have inconsistent distribution and representation, which leads to a modality gap. The fusion network can not compose the semantic information captured from the source image and the modified text sufficiently without erasing the modality gap before fusing. Inspired by the recent advance of Deep InfoMax, we use mutual information maximization the image modality and the text modality (ITDIM) to narrow their modality gap. Considering that the source image and the modified text are completely semantically different, it’s hard to narrow their modality gap by capturing the non-linear statistical dependencies using MI maximization between their representation [22]. Thus, ITDIM maximizes mutual information between the representation of the desired image and that of the modified text which contain partially same semantic information. In the previous work of the image to text retrieval, researchers attempt to obtain a better alignment of distributions of item representations across modalities by adversarial learning. They treat the input query encoder as the ”generator” in GAN [42] and design a modality classifier, which acts as the ”discriminator”. Given an unknown feature projection, this classifier detects the modality of an item as reliably as possible [20, 24]. In essence, these methods are equivalent to maximizing the mutual information between the feature vectors of the different

modalities with the same semantic information. Further, narrowing modality gap by adversarial loss can be viewed as the special cases of MINE [22], the basic version of Deep Infomax. However, it is often insufficient to quantify the dependency between two modalities by estimating mutual information between the two complete representations (i.e., feature vectors), namely global MI maximization. Rather, combining the average MI maximization between the high-level representation and local regions of the low-level representation (e.g., patches rather than the complete representation) [1], namely local MI maximization, can get a better distribution alignment. When the representation is the outputs of different layers of the same encoder, the local MI maximization makes the encoder prefer information that is shared among patches and filter noise specific to local patches. For our task, each modality makes use of different encoders and the semantic information the modified text contains is part of that the desired image contains. If we maximize the local MI between the high-level representation of the modified text and the low-level representation of the desired image, the desired image encoder will discard some image-specific semantic information as noise. Thus, our ITDIM maximizes mutual information between the high-level representation in the desired image encoder and the low-level representation in the modified text encoder. We verify the said analysis through experiments in section 4.5.3.

In this paper, we maximize MI using different  $MI(X; Y)$  objectives, where  $X$  is a low-level feature map, and  $Y$  is a high-level feature vector [1]. Generally speaking, mutual information quantifies the dependence of  $X$  and  $Y$ . We formulate it as follows,

$$I(X; Y) = \int_{\mathcal{X} \times \mathcal{Y}} \log \frac{d\mathbb{P}_{XY}}{d\mathbb{P}_X \otimes \mathbb{P}_Y} d\mathbb{P}_{XY} = \int_{\mathcal{X} \times \mathcal{Y}} \log \frac{d\mathbb{P}_{XY/X}}{d\mathbb{P}_Y} d\mathbb{P}_{XY} \quad (4)$$

where  $\mathbb{P}_{XY}$  is the joint probability distribution, and  $\mathbb{P}_X = \int_{\mathcal{Y}} \mathbb{P}_{XY}$  and  $\mathbb{P}_Y = \int_{\mathcal{X}} \mathbb{P}_{XY}$  are the marginals [22]. In the original Deep InfoMax,  $X$  and  $Y$  are the different-level representations in an encoder and contain the same semantic information. According the Equation (4), the mutual information maximization makes  $Y$  capture the representative information of  $X$  as much as possible. When

we set  $X$  and  $Y$  as the representations of the modified text and the desired image, ITDIM actually guarantees the image representation to hold the semantic information related to the modified text as much as possible.

To make the following section more clear, we give some notations. We define the modified text, the source image, the desired image and their features as  $t$ ,  $s$ ,  $d$ ,  $\phi_t$ ,  $\phi_s$  and  $\phi_d$ , respectively. The fusion feature is denoted by  $\phi_{st}$ . Then we define the input image as  $I \in (s, d)$  and its feature as  $\phi_I \in (\phi_s, \phi_d)$ . In the image encoder, the representation is defined as  $i = I_m(I, \theta_{im})$  where  $I_m$  denotes the image CNN before  $i$  and  $\theta_{im}$  denotes the parameters of this CNN. To obtain the best retrieval accuracy, we set  $i$  to varied layers in terms of different MI objectives and datasets. In the text encoder, we set each representation as  $e = E_m(t, \theta_{tm})$ , which  $E_m$  is the LSTM network with parameters  $\theta_{tm}$ . The text encoder in our method consists of two representations: the output and the parallel connection of the word embeddings.

The key to maximize the MI is to design an appropriate MI estimator. Noise-Contrastive Estimation [58, 59] and Donsker-Varadhan estimators [60] require a large number of negative samples to be competitive and quickly becomes cumbersome with increasing batch size. By contrast, Jensen-Shannon MI estimator [1, 25] performs well using a small quantity of negative samples, so we apply this estimator to our model. Our estimator  $\hat{\mathcal{I}}_{\theta_d, \theta_{tm}, \theta_{im}}^{\text{JSD}}$  for  $(e; i)$ ,  $i = \phi_d$  can be formulated as,

$$\hat{\mathcal{I}}_{\theta_d, \theta_{tm}, \theta_{im}}^{\text{JSD}}(e; i) := \mathbb{E}_{\mathbb{P}_e}[-\text{sp}(-T_{\theta_d, \theta_{tm}, \theta_{im}}(e, i))] - \mathbb{E}_{\mathbb{P}_e \times \mathbb{P}_{e'}}[\text{sp}(T_{\theta_d, \theta_{tm}, \theta_{im}}(e', i))] \quad (5)$$

where  $\mathbb{P}_e$  is the empirical probability distribution of text representation  $e$ ,  $e'$  is the low-level representation sampled from  $\mathbb{P}_{e'} = \mathbb{P}_e$ ,  $\text{sp}(z) = \log(1 + e^z)$  and  $T$  can be concretized as a discriminator function modeled by deep neural network with parameters  $\theta_d$ .

To compute the mutual information between high dimensional representation pairs effectively and sufficiently, we maximize MI by adopting global MI objectives and local MI objective, which maximizes the MI between the com-

plete  $X$  and  $Y$  and estimates the MI between  $Y$  and local regions of  $X$  and respectively. Based on our estimator, we define our global MI and local MI objectives as follows,

$$MI_E^G(e; i) = \max_{\theta_{dg}, \theta_{tm}, \theta_{im}} \hat{\mathcal{I}}_{\theta_{dg}, \theta_{tm}, \theta_{im}}^{\text{JSD}}(e; i) \quad (6)$$

$$MI_E^L(e; i) = \max_{\theta_{dl}, \theta_{tm}, \theta_{im}} \frac{1}{M^2} \sum_{i=1}^{M^2} \hat{\mathcal{I}}_{\theta_{dl}, \theta_{tm}, \theta_{im}}^{\text{JSD}}(e^p; i) \quad (7)$$

where  $\theta_{dg}$  and  $\theta_{dl}$  are the parameters of the discriminators for the global and local MI objectives and  $e^p$  is the  $p$ th patch of the feature map  $e$ .

Besides increasing the dependence between the text modality and the image modality, we also improve the compactness of the image feature by imposing prior matching objective, which makes  $Y$  match a prior distribution. This objective can be formulated as,

$$MI_E^P(i) = \mathbb{E}_{\mathbb{V}_y}[\log \mathcal{D}_{\theta_{dp}}(y)] + \mathbb{E}_{\mathbb{P}_e}[\log(1 - \mathcal{D}_{\theta_{dp}}(i))] \quad (8)$$

where  $y$  denotes a random variable with prior probability distribution  $\mathbb{V}_y$  and  $\theta_{dp}$  is the parameters of the discriminator function  $\mathcal{D}_{\theta_{dp}}$  used in this objective. Finally, we utilize these three objectives together and get the complete objective,

$$L_E = MI_E(e; i) = \alpha MI_E^G(e; i) + \beta MI_E^L(e; i) + \gamma MI_E^P(i) \quad (9)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are the trade-off parameters.

The discriminators in MI objectives vary according to different application scenarios [1]. In our method, we define the discriminators for the global, local and prior matching objectives as Table 1. We set the unit number of most hidden layers to the dimension of each encoder’s output, 512. Since the semantic information in the modified text is a portion of that in the desired image, there may be different text corresponding to the same image. And it’s hard to determine which text-image pairs should have more mutual information. Thus, the ‘fake’ sample  $e'$  in Equation (5) is set as the same low-level feature as the ‘real’ sample  $k$  extracted from another text that is not the description of the desired image.

Objective	Operation	R@size	Activation
Global	Input $\rightarrow$ Linear layer	512	ReLU
	Linear layer	512	ReLU
	Linear layer	1	
Local	Input $\rightarrow 1 \times 1$ conv	512	ReLU
	$1 \times 1$ conv	512	ReLU
	$1 \times 1$	1	
Prior	Input $\rightarrow 1 \times 1$ conv	512	ReLU
	$1 \times 1$ conv	300	ReLU
	$1 \times 1$	1	

Table 1: Network architecture for global DIM, local DIM and prior matching

### 3.2. Distance Metric Learning based on Deep Mutual Information Maximization

The goal of deep metric learning (DML) is to push closer the fusion feature  $\phi_{st}$  and the desired image’s feature  $\phi_d$  while pulling apart the non-similar image’s feature  $\phi_n$ . More precisely, suppose the training minibatch has  $B$  queries, we select one fusion feature  $\phi_{st}^t$  and create a corresponding set  $\mathcal{N}_i$  that consists of a desired image  $\phi_d^t$  and  $K - 1$  non-similar images  $\phi_n^1, \dots, \phi_n^{K-1}$ . We repeat this selection  $M$  times and denote the  $m$ th selection as  $\mathcal{N}_i^m$ . We adopt the following softmax cross-entropy loss,

$$L_T = -\frac{1}{MB} \sum_{t=1}^B \sum_{m=1}^M \log \left\{ \frac{\exp\{\kappa(\phi_{st}^t, \phi_d^t)\}}{\sum_{\phi_a \in \mathcal{N}_i^m} \exp\{\kappa(\phi_{st}^t, \phi_a)\}} \right\} \quad (10)$$

where  $\kappa$  is a similarity kernel and can be implemented as the dot product or negative  $l_2$  distance. If we apply large  $K$  to the above equation, each desired image is contrasted with a lot of other non-similar images. Our model becomes more discriminative and fits faster than that using small  $K$ , but can be more apt to overfitting. Hence, we adopt this model to the dataset which is difficult to converge. In our experience, we use  $K = B$  and  $M = 1$  for Fashion200k and

the Equation (11) can be rewritten as follows,

$$L_T = -\frac{1}{B} \sum_{t=1}^B \log \left\{ \frac{\exp\{\kappa(\phi_{st}^t, \phi_d^t)\}}{\sum_{j=1}^B \exp\{\kappa(\phi_{st}^t, \phi_d^j)\}} \right\}. \quad (11)$$

By contrast,  $K$  can also be set very small. In the extreme case, when we use the smallest value of  $K = 2$ , the loss is the same as the soft triplet loss in the previous literature [61, 62]. The loss function can be formulated as follows,

$$L_T = \frac{1}{MB} \sum_{t=1}^B \sum_{m=1}^M \log\{1 + \exp\{\kappa(\phi_{st}^m, \phi_n^{m,t}) - \kappa(\phi_{st}^m, \phi_d^t)\}\} \quad (12)$$

where  $\phi_n^{m,t}$  denotes the  $t$ th fusion feature in the  $\mathcal{N}_i^m$ . This loss function is applied to the other two datasets, namely MIT-States and CSS.

Compared with the DML in the unimodal scenario [63, 64, 65, 66], the precondition of the similarity learning between different modalities is to learn a common subspace where the items of different modalities can be directly compared to each other. As the inputs of the DML are the fusion feature and the desired image feature, there is also a modality gap existing due to their inconsistent distribution. Compared to the modality gap between the image and the text in the last section, the modality gap between the fusion and the image is smaller because most of the semantic information in the fusion modality comes from the image modality. Hence, the distributions of the features of the image and the fusion are similar to some extent. If we want to get better retrieval performance, we need to improve the similarity until these two distributions are highly consistent. We achieve this goal by maximizing the mutual information between the representations of the fusion and the desired image, which contains the same semantic information. Since TIRG obtains the fusion feature by adding the gating feature and the residual feature, no feature map which contains all semantic information in the fusion network can be used as low-level representation in the local MI objective. If we maximize MI between the high-level layer in the gating or the residual network which contains partial semantic information of the desired image and the low-level layer in the desired image encoder, local MI objective will discard partial semantic information that is unique to desired

image as noise. Thus, we maximize mutual information between the low-level representation in the desired image encoder and the high-level representation in the fusion network. Furthermore, experiments in section 4.5.3 demonstrate that using different layers in the desired image encoder as  $X$  in the global and local MI objectives is much better than using fusion feature  $(\phi_{st})$  as  $X$  in global MI objective. Thus, we set  $Y$  in  $MI(X; Y)$  as the high-level layer in the fusion network and  $X$  as the low-level layer in the desired image’s encoder,  $i = I_m(I, \theta_{im})$ , which is defined in the last section. The mutual information maximization here is between the image modality and the fusion modality, so we call it IFDIM.

The setting of  $X$  and  $Y$  makes IFDIM optimizes the parameters of the whole architecture. We define  $\theta_a$  as the parameters of the entire model and our cross-modal Jensen-Shannon MI estimator can be written as,

$$\widehat{\mathcal{I}}_{\theta_d, \theta_a}^{CJ}(i; \phi_{st}) := \mathbb{E}_{\mathbb{P}_i}[-\text{sp}(-T_{\theta_d, \theta_a}(i, \phi_{st}))] - \mathbb{E}_{\mathbb{P}_i \times \mathbb{P}_{i'}}[\text{sp}(T_{\theta_d, \theta_a}(i', \phi_{st}))] \quad (13)$$

where  $\mathbb{P}_i$  is the empirical probability distribution of  $i$ ,  $\mathbb{P}_{i'} = \mathbb{P}_i$  is the distribution of  $i'$  and  $T$  stands for a discriminator function with parameters  $\theta_d$  as Equation (5). As section 3.1, we increase the dependence between the fusion modality and the image modality by global MI, local MI and prior matching objectives. Since the formulas of these objectives are similar and can be obtained by altering corresponding parameters, we directly provide the complete objective,

$$MI_F(i; \phi_{st}) = \alpha MI_F^G(i; \phi_{st}) + \beta MI_F^L(i; \phi_{st}) + \gamma MI_F^P(\phi_{st}) \quad (14)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are trade-off parameters defined in Equation (5). And the loss function for the cross-modal Deep InfoMax can be represented as  $L_F = MI_F(i; \phi_{st})$ .

Finally, we train our model by the overall loss function defined as,

$$L_{ALL} = \mu(L_E + L_F) + L_T \quad (15)$$

where  $\mu$  is dynamic tradeoff hyperparameters.



## 4. Experiments

This section consists of three parts: 1) introduce the experimental settings; 2) compare our method with the state-of-the-art algorithms on different datasets; 3) provide ablation experiments to study the effect of the ITDIM and the IFDIM in our model.

### 4.1. Experimental Settings

We compare our method with TIRG [14] and seven benchmarks mentioned in section 2.1 on three datasets: Fashion200k [26], MIT-States [27] and CSS [14]. Our main metric for retrieval is recall at rank  $k$  ( $R@k$ ), computed as the percentage of the text queries where (at least 1) desired or correct labeled image is within the top  $K$  retrieval images. In order to get stable retrieval results, we repeat each experiment 5 times, and both mean and standard deviation are reported. We use PyTorch in our experiments. For all datasets, the low-level representation in  $MI(X;Y)$  objectives used in the image encoder are set as the last convolutional layer for its better performance. By default, training is run for 160k iterations with a start learning rate 0.01. We will release the code to the public. The weights  $\alpha$ ,  $\beta$  and  $\gamma$  are set as 0.5, 1 and 0.1. We set  $\mu = L_T/15(L_E + L_F)$  with initial value 0.001 and update it every 10k iterations. We apply our method TIRG-DIM<sub>A</sub> to Fashion200k and MIT-States and TIRG-DIM<sub>C</sub> to CSS in terms of their modified text’s attribute.

### 4.2. Fashion200k

Fashion200k is a widely-used dataset in the filed of cross-modal image retrieval. It is composed of 200k images of fashion products and each image has a compact attribute-like description (such as mini and short dress or knee length skirt). Following the previous work [26], queries are generated as follows: the query images and its desired images have one word difference in their descriptions, and the modified text is this different word. We adopt the same training split as TIRG [14] and generate queries on the fly for training. We randomly sample 10 validation sets of 3167 test queries and report the mean.



Figure 3: Qualitative results of image retrieval with modified text on Fashion200k. blue/green boxes: source/desired images.



Figure 4: Qualitative results of image retrieval with modified text on MIT-states. blue/green boxes: source/desired images.

Method	R@1	R@10	R@50
[26]	6.3	19.9	38.3
Image only	3.5	22.7	43.7
Text only	1.0	12.3	21.8
Concatenation	11.9 $\pm$ 1.0	39.7 $\pm$ 1.0	62.6 $\pm$ 0.7
Show and Tell	12.3 $\pm$ 1.1	40.2 $\pm$ 1.7	61.8 $\pm$ 0.9
Param Hashing	12.2 $\pm$ 1.1	40.0 $\pm$ 1.1	61.7 $\pm$ 0.8
Relationship	13.0 $\pm$ 0.6	40.5 $\pm$ 0.7	62.4 $\pm$ 0.6
Film	12.9 $\pm$ 0.7	39.5 $\pm$ 2.1	61.9 $\pm$ 1.9
TIRG	<u>14.1<math>\pm</math>0.6</u>	<u>42.5<math>\pm</math>0.7</u>	<u>63.8<math>\pm</math>0.8</u>
TIRG-DIM <sub>A</sub>	<b>17.4<math>\pm</math>0.3</b>	<b>43.4<math>\pm</math>0.4</b>	<b>64.5<math>\pm</math>0.6</b>

Table 2: Retrieval performance on Fashion200k. The best result is in bold and the second best in underline.

Figure 3 illustrates some qualitative results and Table 2 shows the retrieval accuracy on this dataset. From the results we have the following observations: 1) our method outperforms all the other approaches with a large margin, especially the R@1 performance, which has a more than 23 percent increase over the best competitor; 2) The standard deviations of the proposed method are smaller than others. These observations demonstrate that we can get more accurate and stable retrieval performance by improving the quality of each input’s feature and the fusion feature using mutual information maximization.

#### 4.3. MIT-States

MIT-States has 63440 images, and each image is described by an object/noun word and a state/adjective word (such as wide belt or tiny island). In total, this dataset contains 245 nouns and 115 adjectives and each individual noun is only modified by 9 adjectives it affords.

For image retrieval, we create the query image and desired image by sampling pairs of images with the same object labels and different state labels. The state of the desired image is considered as the modified text. Therefore, our method

Method	R@1	R@5	R@10
Image only	3.3	12.8	20.9
Text only	7.4	21.5	32.7
Concatenation	11.8 $\pm$ 0.2	30.8 $\pm$ 0.2	42.1 $\pm$ 0.3
Show and Tell	11.9 $\pm$ 0.1	31.0 $\pm$ 0.5	42.0 $\pm$ 0.8
Att. as Operator	8.8 $\pm$ 0.1	27.3 $\pm$ 0.3	39.1 $\pm$ 0.3
Relationship	<u>12.3<math>\pm</math>0.5</u>	<u>31.9<math>\pm</math>0.7</u>	42.9 $\pm$ 0.9
Film	10.1 $\pm$ 0.3	27.7 $\pm$ 0.7	38.3 $\pm$ 0.7
TIRG	12.2 $\pm$ 0.4	<u>31.9<math>\pm</math>0.3</u>	<u>43.1<math>\pm</math>0.3</u>
TIRG-DIM <sub>A</sub>	<b>14.1<math>\pm</math>0.3</b>	<b>33.8<math>\pm</math>0.5</b>	<b>45.0<math>\pm</math>0.5</b>

Table 3: Retrieval performance on MIT-States. The best result is in bold and the second best in underline.

is to retrieve image which possesses the same object but new state compared with the query image. In the experiments, we select 80 nouns for training, and the others are adopted for testing. Based on these settings, models are trained by different state/adjective (modified text) and tested using unseen objects.

A number of qualitative results are shown in Figure 4 and the quantitative retrieval results can be seen from Table 3. Obviously, our proposed method obtains the highest retrieval accuracy at different R@k on this dataset. More specifically, we achieve 15% and 6% improvement on R@1 and R@10 respectively compared with the second best algorithm, namely Relationship [30]. Because the same object with varied states can look extremely different, the modification text becomes more significant. Therefore, the “Text only” baseline overcomes “Image only”.

#### 4.4. CSS dataset

CSS consists of 32k synthesized images in a 3-by-3 grid scene which are generated by CLEVR toolkit. Objects in the images are rendered with different color, shape and size occupy. Each image comes in a simple 2D blobs version and 3D version, and we utilize the second one in this paper. There are 16k queries for

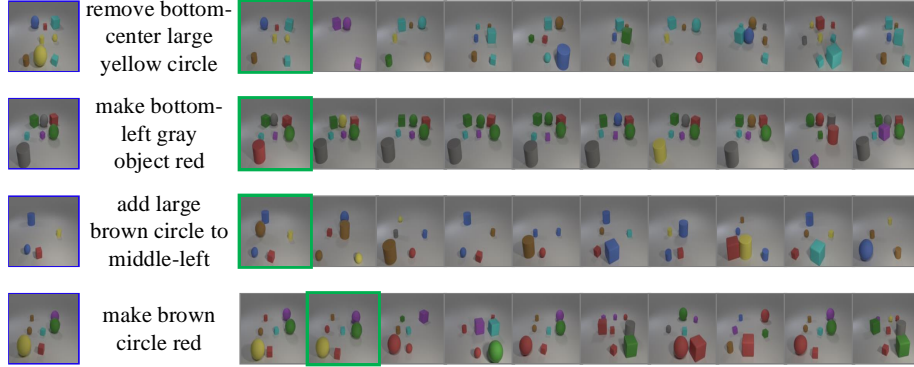


Figure 5: Qualitative results of image retrieval with modified text on CSS. blue/green boxes: source/desired images.

training and 16k queries for testing in this dataset. Each query is composed of a source image, a modified text and a desired image (Figure 1). The modification text has three templates: adding, removing or changing object attributes, such as "add small green rectangle to top-right", "remove bottom-center small red circle" or "make bottom-left large green object gray". We provide a stronger test of generation by making certain object shape and color combinations only appear in training and not in testing, and vice versa.

We can find the quantitative and qualitative results from Figure 5 and Table 4 respectively. All the methods except the Image Only and the Text Only approaches get much higher retrieval accuracy on this dataset than on the other two. We believe this is because the image queries are simple and the text queries contain more information. Compared to the second best method TIRG, TIRG-DIM<sub>A</sub> improves retrieval accuracy by 3.3, 4.9 and 3.0 percentage points on R@1, R@5 and R@10 score.

#### 4.5. Ablation Studies

In this section, we first provide the R@1 accuracy of various ablation studies to gain insight into which part of our method matters the most. The results are in Table 5. Then we offer the loss values and the visualization of distribution by Figure 6 and Figure 7.

Method	R@1	R@5	R@10
Image only	6.3	29.3	54.0
Text only	0.1	0.5	0.8
Concatenation	60.6 $\pm$ 0.5	88.2 $\pm$ 0.4	92.8 $\pm$ 0.4
Show and Tell	33.0 $\pm$ 3.2	75.0 $\pm$ 1.3	83.0 $\pm$ 0.9
Para.Hasing	60.5 $\pm$ 1.9	88.1 $\pm$ 0.8	92.9 $\pm$ 0.6
Relationship	62.1 $\pm$ 1.2	89.1 $\pm$ 0.4	93.5 $\pm$ 0.7
Film	65.6 $\pm$ 0.5	89.7 $\pm$ 0.6	94.1 $\pm$ 0.5
TIRG	<u>73.7</u> $\pm$ 0.4	<u>90.7</u> $\pm$ 0.4	<u>94.6</u> $\pm$ 0.4
TIRG-DIM <sub>A</sub>	<b>77.0</b> $\pm$ 0.2	<b>95.6</b> $\pm$ 0.4	<b>97.6</b> $\pm$ 0.3

Table 4: Retrieval performance on CSS. The best result is in bold and the second best in underline.

#### 4.5.1. Effect of ITDIM

We study the effect of the deep mutual information maximization between low-level representation in the text encoder and the high-level representation in the desired image encoder (ITDIM) in both TIRG<sub>A</sub> and TIRG<sub>C</sub>. From the results in Table 5, we can see that the performance of the two models has a remarkable improvement by adding ITDIM into these models. It demonstrates that the ITDIM can help to get a better alignment of distributions of item representations between the modified text and the image, though the semantic information in the text modality is much less than that in the image modality.

#### 4.5.2. Effect of IFDIM

Comparing the results obtained with deep mutual information maximization between the low-level representation in the desired image encoder and the high-level representation in the fusion network (IFDIM) and that with ITDIM, we can see the models based on IFDIM leads to more considerable improvement. We believe this is because the representations of these two modalities all contain rich semantic information. The performance of DIM in our model is partly related to the quantity of the semantic information contained in its two inputs.

#### 4.5.3. Effect of other Deep InfoMaxs

In order to give a more detailed comparison between different deep InfoMaxs, we also provide experiments based on deep mutual information maximization using other low-level and high-level representation in Table 5.

To obtain better fusion features, we attempt to make use of  $\text{DIM}_{\text{TextSour}}$ , which maximizes mutual information between the low-level representation in the text encoder and the high-level representation in the source image encoder, and  $\text{DIM}_{\text{SourText}}$ , which swaps the positions of the representation in MI objectives. As the low-level and high-level representation in these InfoMaxs are semantically different, it’s harmful to retrieval to force Deep InfoMax to capture non-linear statistical dependencies between two semantically independent modalities. We also employ  $\text{DIM}_{\text{SourDes}}$ ,  $\text{DIM}_{\text{DesSour}}$  and  $\text{DIM}_{\text{DesText}}$  (Des in the subscripts represents the representation in the the desired image encoder), whose low-level representation contain partially different semantic information compared to their high-level representation. In these InfoMaxs, local mutual information objectives discard part of semantic information that is unique to the low-level representation as mentioned in section 3.1, which affect the retrieval results. Moreover, we try to optimize the fusion network by aligning the distribution between the text representation and the fusion feature by  $\text{DIM}_{\text{TextFus}}$  and narrowing the modality gap between the source image and the fusion feature by  $\text{DIM}_{\text{SourFus}}$ . The experimental results show that these two InfoMaxs marginally improve the performance. We think this is because the effect of mutual information maximization between the low-level representation and the high-level representation of the same deep neural network is limited in fully supervised learning. However, ITDIM can significantly improve the retrieval accuracy by estimating mutual information between the representation from different encoders.

For distance metric learning, learning a common space between fusion feature and the desired image feature is the key to guarantee the items of different modalities can be directly compared to each other. Apart from IFDIM, we

exploit Deep InfoMax between the low-level representation in the fusion network and the high-level representation in the desired image encoder. Considering that fusion feature in TIRG is the weighted sum of the residual feature and the gating feature, no feature map which contains all semantic information in the fusion network can be used as low-level representation in the local MI objective. Here we conduct experiments based on  $\text{DIM}_{FusDes}$ ,  $\text{DIM}_{ResiDes}$  and  $\text{DIM}_{GatingDes}$ , which give up using the local MI objective, apply feature map in the residual network and feature map in the gating work to local MI objective respectively. Although these Deep InfoMaxs can't compete with ITDIM and IFDIM, they overcome the others.

#### 4.5.4. Qualitative Analyses of the Effects of ITDIM and IFDIM

Apart from the quantitative analysis of the DIM effect using retrieval accuracy, we also provide qualitative analysis based on the trend of the loss and the visualization of distribution. From Figure 6, we can find that the models' loss trends using different DIMs vary each other. Overall, the models with DIM perform better than that without DIM. The larger the number of iterations, the greater the difference becomes. In detail, TIRG-DIM is superior to the methods with ITDIM or IFDIM. It shows that aligning the distribution of all three modalities performs better than aligning two of them, which is in line with common sense. The TIRG-IFDIM is a little bit better than TIRG-ITDIM in terms of the training loss. This is because the semantic information of the inputs in IFDIM is rich and identical, while the text modality in ITDIM is semantically deficient.

We also visualize the distributions of the features of the desired image and the fusion of source image and the text on the MIT-states dataset using t-SNE tool (1280 sample points for each modality) by Figure 7. An ideal distribution alignment across modalities should make their features project into a compact common subspace like Figure 7(d), and vice versa. Figure 7(a), Figure 7(b), Figure 7(c) and Figure 7(d) represent the distribution of the features learned by TIRG, TIRG-ITDIM, TIRG-IFDIM and TIRG-DIM (ITDIM+IFDIM), re-



spectively. The alignment of the distributions of these two features realized by the models with mutual information maximization is much better than that realized by the models without mutual information maximization. Obviously, TIRG-DIM reaches the best alignment of distributions of item representations across modalities. It should be noted that the model based on IFDIM learns a better common subspace for the image modality and the fusion modality compared to the model based on ITDIM, which corresponds to trends of the loss in Figure 6.

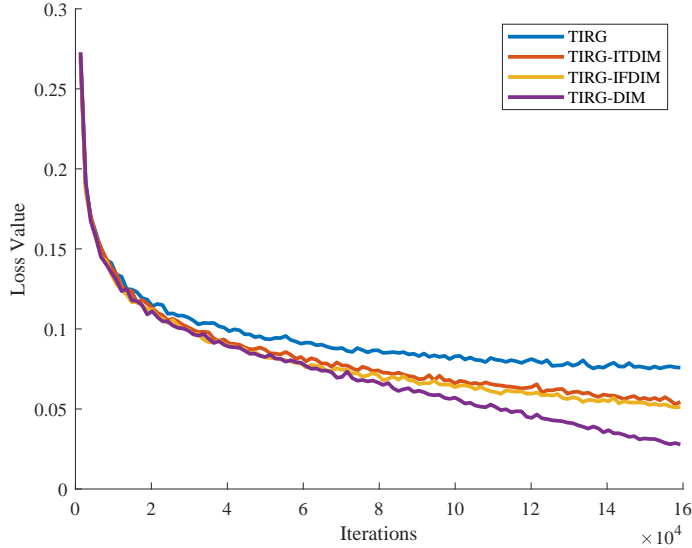
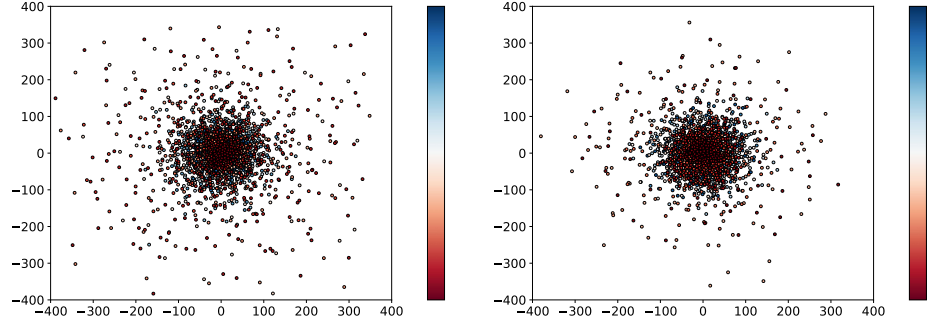


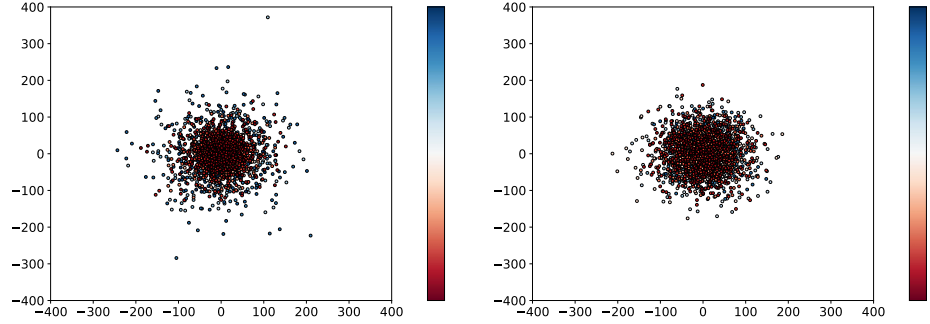
Figure 6: The training loss of the distance metric learning using different models on MIT-States dataset.

## 5. Conclusion

In this paper, we have proposed a new method for cross-modal image retrieval based on the contrastive self-supervised learning method Deep InfoMax [22, 1]. Our approach makes retrieval more accurate by aligning the feature distributions of text, image, and their fusion. We maximize the MI between semantically different representations of the image modality and the text modality



(a) The cross-modal invariance preserved with- (b) The cross-modal invariance preserved with  
out DIM ITDIM



(c) The cross-modal invariance preserved with (d) The cross-modal invariance preserved with  
IFDIM DIM

Figure 7: The cross-modal invariance preserved. We visualise the distribution of the fusion modality and the image modality using the dots of different colors. The red and green solid dots represent the sampled fusion features and desired image features, respectively. We specify different lightness for the dots of the same color (as shown in the colorbar) to make them easier to distinguish.

to project the features of these two modalities into a common subspace. Moreover, our method gets a precise alignment of distribution of the image modality and the fusion modality by maximizing the MI between the semantically identical representations in the desired image encoder and the fusion network. The proposed method gives an improved performance on three benchmark datasets. In the future, we would like to apply our work to other areas of cross-modal retrieval.

## 6. Acknowledgments

This work is supported by Alibaba-Zhejiang University Joint Institute of Frontier Technologies, The National Key R&D Program of China (No. 2018YFC2002603, 2018YFB1403202), Zhejiang Provincial Natural Science Foundation of China (No. LZ13F020001), the National Natural Science Foundation of China (No. 61972349, 61173185, 61173186) and the National Key Technology R&D Program of China (No. 2012BAI34B01, 2014BAK15B02).

## References

- [1] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, arXiv preprint arXiv:1808.06670.
- [2] S. Chopra, R. Hadsell, Y. LeCun, et al., Learning a similarity metric discriminatively, with application to face verification, in: CVPR (1), 2005, pp. 539–546.
- [3] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: European conference on computer vision, Springer, 2016, pp. 241–257.
- [4] F. Cakir, K. He, X. Xia, B. Kulis, S. Sclaroff, Deep metric learning to rank, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1861–1870.

- [5] W. Cao, Q. Lin, Z. He, Z. He, Hybrid representation learning for cross-modal retrieval, *Neurocomputing* 345 (2019) 45–57.
- [6] X. Tian, X. Zhou, W. W. Ng, J. Li, H. Wang, Bootstrap dual complementary hashing with semi-supervised re-ranking for image retrieval, *Neurocomputing* 379 (2020) 103–116.
- [7] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.
- [8] Z. Hu, Y. Luo, J. Lin, Y. Yan, J. Chen, Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching, in: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, AAAI Press, 2019, pp. 789–795.
- [9] P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: learning to retrieve badly drawn bunnies, *ACM Transactions on Graphics (TOG)* 35 (4) (2016) 1–12.
- [10] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, Y.-Z. Song, Generalising fine-grained sketch-based image retrieval, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 677–686.
- [11] T.-Y. Lin, Y. Cui, S. Belongie, J. Hays, Learning deep representations for ground-to-aerial geolocalization, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5007–5015.
- [12] S. Hu, M. Feng, R. M. Nguyen, G. Hee Lee, Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267.
- [13] B. Zhao, J. Feng, X. Wu, S. Yan, Memory-augmented attribute manipulation networks for interactive fashion search, in: *Proceedings of the IEEE*

- Conference on Computer Vision and Pattern Recognition, 2017, pp. 1520–1528.
- [14] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, J. Hays, Composing text and image for image retrieval-an empirical odyssey, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6439–6448.
  - [15] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, R. Feris, Dialog-based interactive image retrieval, in: Advances in Neural Information Processing Systems, 2018, pp. 678–688.
  - [16] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences, Transactions of the Association for Computational Linguistics 2 (2014) 207–218.
  - [17] T. Nagarajan, K. Grauman, Attributes as operators: factorizing unseen attribute-object compositions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 169–185.
  - [18] H. Noh, P. Hongsuck Seo, B. Han, Image question answering using convolutional neural network with dynamic parameter prediction, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 30–38.
  - [19] E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: Visual reasoning with a general conditioning layer, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
  - [20] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H. T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 154–162.

- [21] J. B. Kinney, G. S. Atwal, Equitability, mutual information, and the maximal information coefficient, *Proceedings of the National Academy of Sciences* 111 (9) (2014) 3354–3359.
- [22] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, R. D. Hjelm, Mine: mutual information neural estimation, *arXiv preprint arXiv:1801.04062*.
- [23] P. Bachman, R. D. Hjelm, W. Buchwalter, Learning representations by maximizing mutual information across views, in: *Advances in Neural Information Processing Systems*, 2019, pp. 15535–15545.
- [24] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, S. C. Hoi, Learning cross-modal embeddings with adversarial networks for cooking recipes and food images, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11572–11581.
- [25] S. Nowozin, B. Cseke, R. Tomioka, f-gan: Training generative neural samplers using variational divergence minimization, in: *Advances in neural information processing systems*, 2016, pp. 271–279.
- [26] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, L. S. Davis, Automatic spatially-aware fashion concept discovery, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1463–1471.
- [27] P. Isola, J. J. Lim, E. H. Adelson, Discovering states and transformations in image collections, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1383–1391.
- [28] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: Visual question answering, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [29] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

- [30] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: *Advances in neural information processing systems*, 2017, pp. 4967–4976.
- [31] S. Becker, An information-theoretic unsupervised learning algorithm for neural networks, University of Toronto, 1992.
- [32] S. Becker, Mutual information maximization: models of cortical self-organization, *Network: Computation in neural systems* 7 (1) (1996) 7–31.
- [33] L. Wiskott, T. J. Sejnowski, Slow feature analysis: Unsupervised learning of invariances, *Neural computation* 14 (4) (2002) 715–770.
- [34] A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural computation* 7 (6) (1995) 1129–1159.
- [35] R. Linsker, Self-organization in a perceptual network, *Computer* 21 (3) (1988) 105–117.
- [36] L. Paninski, Estimation of entropy and mutual information, *Neural computation* 15 (6) (2003) 1191–1253.
- [37] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, *arXiv preprint arXiv:1906.05849*.
- [38] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, *arXiv preprint arXiv:1807.03748*.
- [39] O. J. Hénaff, A. Razavi, C. Doersch, S. Eslami, A. v. d. Oord, Data-efficient image recognition with contrastive predictive coding, *arXiv preprint arXiv:1905.09272*.
- [40] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, *arXiv preprint arXiv:1911.05722*.

- [41] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, arXiv preprint arXiv:2002.05709.
- [42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
- [43] M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, arXiv preprint arXiv:1701.04862.
- [44] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv preprint arXiv:1701.07875.
- [45] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, B. Dolan, Generating informative and diverse conversational responses via adversarial information maximization, in: Advances in Neural Information Processing Systems, 2018, pp. 1810–1820.
- [46] D. Qian, W. K. Cheung, Enhancing variational autoencoders with mutual information neural estimation for text generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 4038–4048.
- [47] A. D. McCarthy, X. Li, J. Gu, N. Dong, Improved variational neural machine translation by promoting mutual information, arXiv preprint arXiv:1909.09237.
- [48] L. Kong, C. d. M. d’Autume, W. Ling, L. Yu, Z. Dai, D. Yogatama, A mutual information maximization perspective of language representation learning, arXiv preprint arXiv:1910.08350.
- [49] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, M. Lucic, On mutual information maximization for representation learning, arXiv preprint arXiv:1907.13625.



- [50] L. Wen, Y. Zhou, L. He, M. Zhou, Z. Xu, Mutual information gradient estimation for representation learning, arXiv preprint arXiv:2005.01123.
- [51] N. Sayed, B. Brattoli, B. Ommer, Cross and learn: Cross-modal self-supervision, in: German Conference on Pattern Recognition, Springer, 2018, pp. 228–243.
- [52] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [53] W. Guo, H. Huang, X. Kong, R. He, Learning disentangled representation for cross-modal retrieval with deep mutual information estimation, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 1712–1720.
- [54] R. Vemulapalli, H. Van Nguyen, S. K. Zhou, Deep networks and mutual information maximization for cross-modal medical image synthesis, in: Deep Learning for Medical Image Analysis, Elsevier, 2017, pp. 381–403.
- [55] R. Krishna, M. Bernstein, L. Fei-Fei, Information maximizing visual question generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2008–2018.
- [56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [57] R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information, arXiv preprint arXiv:1703.00810.
- [58] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 297–304.

- [59] M. U. Gutmann, A. Hyvärinen, Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, *Journal of Machine Learning Research* 13 (Feb) (2012) 307–361.
- [60] M. D. Donsker, S. S. Varadhan, Asymptotic evaluation of certain markov process expectations for large time. iv, *Communications on Pure and Applied Mathematics* 36 (2) (1983) 183–212.
- [61] N. N. Vo, J. Hays, Localizing and orienting street views using overhead imagery, in: *European Conference on Computer Vision*, Springer, 2016, pp. 494–509.
- [62] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, *arXiv preprint arXiv:1703.07737*.
- [63] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification., *Journal of Machine Learning Research* 10 (2).
- [64] C. Gu, J. Bu, K. Shi, Z. Yu, B. Wang, L. Li, Local metric learning based on anchor points for multimedia analysis, in: *2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2019, pp. 1366–1371.
- [65] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4004–4012.
- [66] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5022–5030.

Method	Fashion	MIT-State	CSS
TIRG <sub>A</sub>	14.1 $\pm$ 0.6	12.2 $\pm$ 0.4	71.2 $\pm$ 0.4
TIRG <sub>A</sub> + DIM <sub>TextSour</sub>	13.6 $\pm$ 0.6	11.6 $\pm$ 0.5	70.5 $\pm$ 0.6
TIRG <sub>A</sub> + DIM <sub>SourText</sub>	13.7 $\pm$ 0.4	11.7 $\pm$ 0.6	70.7 $\pm$ 0.5
TIRG <sub>A</sub> + DIM <sub>SourDes</sub>	13.5 $\pm$ 0.5	11.4 $\pm$ 0.4	70.4 $\pm$ 0.5
TIRG <sub>A</sub> + DIM <sub>DesSour</sub>	13.3 $\pm$ 0.7	11.3 $\pm$ 0.5	70.2 $\pm$ 0.4
TIRG <sub>A</sub> + DIM <sub>DesText</sub>	13.1 $\pm$ 0.6	11.2 $\pm$ 0.4	70.1 $\pm$ 0.5
TIRG <sub>A</sub> + DIM <sub>TextFus</sub>	14.2 $\pm$ 0.6	12.3 $\pm$ 0.4	71.4 $\pm$ 0.3
TIRG <sub>A</sub> + DIM <sub>SourFus</sub>	14.3 $\pm$ 0.5	12.4 $\pm$ 0.3	71.3 $\pm$ 0.2
TIRG <sub>A</sub> + DIM <sub>FusDes</sub>	14.5 $\pm$ 0.4	12.5 $\pm$ 0.3	71.6 $\pm$ 0.3
TIRG <sub>A</sub> + DIM <sub>ResiDes</sub>	14.7 $\pm$ 0.5	12.6 $\pm$ 0.4	71.7 $\pm$ 0.3
TIRG <sub>A</sub> + DIM <sub>GatingDes</sub>	14.8 $\pm$ 0.4	12.6 $\pm$ 0.3	71.8 $\pm$ 0.3
TIRG <sub>A</sub> + ITDIM	15.4 $\pm$ 0.4	12.7 $\pm$ 0.3	72.1 $\pm$ 0.2
TIRG <sub>A</sub> + IFDIM	16.5 $\pm$ 0.3	13.7 $\pm$ 0.2	73.2 $\pm$ 0.3
TIRG-DIM <sub>A</sub>	17.4 $\pm$ 0.3	14.1 $\pm$ 0.3	73.8 $\pm$ 0.2
TIRG <sub>C</sub>	12.4 $\pm$ 0.5	10.3 $\pm$ 0.5	73.7 $\pm$ 0.4
TIRG <sub>C</sub> + DIM <sub>TextSour</sub>	11.8 $\pm$ 0.6	9.9 $\pm$ 0.5	73.1 $\pm$ 0.5
TIRG <sub>C</sub> + DIM <sub>SourText</sub>	12.0 $\pm$ 0.5	10.0 $\pm$ 0.4	73.3 $\pm$ 0.6
TIRG <sub>C</sub> + DIM <sub>SourDes</sub>	11.6 $\pm$ 0.5	9.8 $\pm$ 0.5	73.1 $\pm$ 0.5
TIRG <sub>C</sub> + DIM <sub>DesSour</sub>	11.5 $\pm$ 0.6	9.6 $\pm$ 0.4	72.9 $\pm$ 0.5
TIRG <sub>C</sub> + DIM <sub>DesText</sub>	11.4 $\pm$ 0.5	9.5 $\pm$ 0.4	72.7 $\pm$ 0.5
TIRG <sub>C</sub> + DIM <sub>TextFus</sub>	12.6 $\pm$ 0.4	10.4 $\pm$ 0.4	74.0 $\pm$ 0.4
TIRG <sub>C</sub> + DIM <sub>SourFus</sub>	12.6 $\pm$ 0.3	10.5 $\pm$ 0.5	73.9 $\pm$ 0.3
TIRG <sub>C</sub> + DIM <sub>FusDes</sub>	12.7 $\pm$ 0.4	10.7 $\pm$ 0.5	74.1 $\pm$ 0.3
TIRG <sub>C</sub> + DIM <sub>ResiDes</sub>	12.9 $\pm$ 0.3	10.8 $\pm$ 0.4	74.3 $\pm$ 0.3
TIRG <sub>C</sub> + DIM <sub>GatingDes</sub>	13.0 $\pm$ 0.3	11.0 $\pm$ 0.3	74.5 $\pm$ 0.3
TIRG <sub>A</sub> + ITDIM	15.4 $\pm$ 0.4	12.7 $\pm$ 0.3	72.1 $\pm$ 0.2
TIRG <sub>C</sub> + IFDIM	13.9 $\pm$ 0.3	12.3 $\pm$ 0.2	76.5 $\pm$ 0.3
TIRG-DIM <sub>C</sub>	14.8 $\pm$ 0.2	12.9 $\pm$ 0.1	77.0 $\pm$ 0.2
Our Full Model	<b>17.4<math>\pm</math>0.3</b>	<b>14.1<math>\pm</math>0.3</b>	<b>77.0<math>\pm</math>0.2</b>

Table 5: Retrieval performance (R@1) of ablation studies