Boosting Cross-Modal Retrieval With MVSE++ and Reciprocal Neighbors

WEI WEI, (Senior Member, IEEE), MENGMENG JIANG, XIANGNAN ZHANG, HENG LIU, AND CHUNNA TIAN, (Member, IEEE)

1Research and Development Institute, Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
2School of Computer Science, Northwestern Polytechnical University, Xi’an 710071, China
3School of Electronic Engineering, Xidian University, Xi’an 710071, China

Corresponding author: Chunna Tian (chnatian@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671385, Grant 61571354; and in part by the Science, Technology, and Innovation Commission of Shenzhen Municipality under Grant JCYJ20190806160210899.

ABSTRACT In this paper, we propose to boost the cross-modal retrieval through mutually aligning images and captions on the aspects of both features and relationships. First, we propose a multi-feature based visual-semantic embedding (MVSE++) space to retrieve the candidates in another modality, which provides a more comprehensive representation of the visual content of objects and scene context in images. Thus, we have more potential to find a more accurate and detailed caption for the image. However, captioning concentrates the image contents by semantic description. The cross-modal neighboring relationships start from the visual and semantic sides are asymmetric. To retrieve a better cross-modal neighbor, we propose to re-rank the initially retrieved candidates according to the k nearest reciprocal neighbors in MVSE++ space. The method is evaluated on the benchmark datasets of MSCOCO and Flickr30K with standard metrics. We achieve higher accuracy in caption retrieval and image retrieval at both R@1 and R@10.

INDEX TERMS Cross-modal retrieval, visual-semantic embedding, scene context, reciprocal neighbors, re-ranking method.

I. INTRODUCTION

The task of image-caption retrieval aims at finding corresponding sentences given an image query or retrieving images with a sentence query. Although image-caption retrieval has wide applications in semantic searching, it is still challenging. It starts from the single-direction retrieval in visual space. Similar images to the query are first retrieved from the training set based on their similarities with the query in the visual space. Then, the captions of the retrieved images are re-ranked and transferred to the query image [1]–[3]. In IM2TEXT model [1], the similarity of candidate images to the query is measured by Gist and Tiny Image descriptors. Then, the candidates are re-ranked based on evaluating the image content, such as objects, stuff, people and so on. However, Mason and Charniak [2] only take the textual information into account in re-ranking. The final retrieved description is determined by extractive summarization techniques [4]. Kuznetsova et al. [3] report the phrase-based approach to synthesize the output. Given a query image, they identify its content elements using classifiers/detectors and then retrieve phrases referring to the content elements from the database. Finally, the retrieved phrases are combined into a description considering the word order and redundancy etc.

In bi-directional retrievals, most approaches share the core idea of building a common multimodal embedding space for the visual and textual data based on a training set of image-description pairs. Then, the query is retrieved in the multimodal space [5]. Therefore, the most critical issue is to build a joint embedding space, which aligns image and text very well. Recently, a mass of methods on joint visual-semantic embedding are proposed, which can be divided into two main categories. Methods in the first category are based on canonical correlation analysis (CCA) [6], including the normalized CCA [7] and kernel CCA [8]. Kernel CCA uses the kernel trick to produce nonlinear, non-parametric projections from the visual and semantic features into the cross-modal space. Recently, CCA is introduced into deep learning framework [9], [10]. However, a drawback of CCA is that it requires loading all data into memory to compute the covariance matrix.
Thus, the consumption of memory is heavy. Another kind of methods are based on ranking loss. For examples, some deep learning methods, such as restricted Boltzmann machines and autoencoders [11]–[15], have been applied to jointly embed the image and text. VSE [16] embeds the deep visual features and deep semantic features into a cross-modal space based on a bi-direction ranking loss. Based on VSE, VSE++ [17] improves the loss function by considering the influence of hard negative samples, which has achieved the state-of-the-art performance. However, the visual features of VSE++ are extracted by the ImageNet based CNN, which ignores the scene context in images. Therefore, we concatenate the object and scene features to fully represent the object and scene-context in images. As a result, we propose to retrieve the candidates in another modality in a multi-feature based VSE++ (MVSE++) space. It provides a more comprehensive representation of the visual content of objects and scene context in images. Thus, we achieve a better visual-semantic alignment on the aspects of objects and scene context. However, captioning concentrates the image contents by semantic description. The cross-modal neighboring relationships start from the visual and semantic sides are asymmetric. To retrieve a better cross-modal neighbor, we propose to re-rank the initially retrieved candidates according to the $k$ nearest reciprocal neighbors in MVSE++ space.

To summarize, the contributions of this paper are:

- We propose to boost the cross-modal retrieval through mutually aligning images and captions on the aspects of both features and relationships. It reduces the retrieval noise induced by mis-alignment, high dimensional vector and asymmetric cross-modal neighboring relationship.
- We propose multiple visual features based embedding method, which provides a more comprehensive representation of the visual content of objects and scene context in images. Therefore, it provides us a better initial retrieval.
- We define the cross-modal $k$-reciprocal nearest neighbors to re-rank the initially retrieved results, which achieves more accurate results through symmetrizing nearest neighborhood relationships in the cross-modal space.
- Experimental results show that the proposed method achieves advanced image-caption retrieval performance on MSCOCO and Flickr30K datasets. To the best of our knowledge, we achieve the highest accuracy in caption retrieval and image retrieval at both R@1 and R@10.

II. RELATED WORK

Our framework involves the ideas of joint deep embedding and re-ranking. Below, we give a concise overview of relevant prior work on both topics.

A. DEEP VISUAL-SEMANTIC EMBEDDING

CCA is introduced to the deep learning framework for visual-semantic embedding to learn nonlinear predictions and improve its adaptability to large training sets [9], [10]. Considering that stochastic gradient descent (SGD) used in [9], [10] cannot guarantee a good solution to the generalized eigenvalue problem of CCA, Wang et al. [18] propose a network sharing a similar two-branch architecture with deep CCA models. Restricted Boltzmann machines and autoencoders [11], [12] are also incorporated into the learning of a multimodal space. Recently, the recurrent model achieves big progress in image-text embedding [16], [19]–[22]. Among the ranking loss based joint embedding, WSABIE [23] and DeVISE [24] learn linear transformations of visual and textual features into a shared space using a single-directional ranking loss, which applies a margin-based penalty to hard incorrect annotations. However, the single-directional ranking loss may lead to incorrect matches in the opposite direction. To tackle this problem, some works [16], [17], [25]–[28] adopt the bi-directional ranking loss. Attention mechanism is also applied to model the joint embedding at image-text tasks [29]–[31], where they selectively focus on a subset of words and image regions to compute the similarity. Ba et al. [32] propose a recurrent attention model, which attends to some label-relevant image regions of an image for multiple objects recognition. In [33], a multimodal context-modulated attention mechanism is proposed to compute the similarities between images and captions. Nam et al. [34] use an attention mechanism on both images and captions to capture fine-grained interplay between vision and language. Adversarial learning provides a novel perspective for cross-modal retrieval. GXN [35] incorporates generative processes into the cross-modal feature embedding, which is based on GAN (Generative Adversarial Networks) and reinforcement learning. However, all of these works use the object-related visual features, while ignore the scene-context information. Thus, we propose to use multiple visual features which contain the scene information besides object information, which is extracted through bi-CNNs. We use gated recurrent unit (GRU) [36] to represent text. Finally, a bi-directional ranking loss is used to optimize the embedding.

B. RE-RANKING METHODS

Due to the high dimensionality of feature vectors, it is hard for similarity measure to completely distinguish the relevant vectors from the irrelevant ones, which means that there are still some noises in the retrieved results. Thus, for retrieval tasks, even an initial ranking list is obtained, re-ranking is still necessary to guarantee the relevant images have higher ranks. Besides, re-ranking does not need additional training samples and is applicable to any initial ranking results. Recently, re-ranking is popular in instance retrieval [37]–[40]. Many works use $k$-nn [41], [42] to measure the similarity of the probe to address the re-ranking problem. Arandjelovic and Zisserman [43] propose a discriminative query expansion (DQE) us a linear SVM to obtain a weight vector, which emphasize the negative samples that are far away from the query image. Chum et al. [41] propose a new query vector, which is obtained by averaging the feature vectors of the
FIGURE 1. The framework of the proposed method.

returned top-$k$ results, and then use it to re-query the database. Shen et al. [39] take the $k$-nn of the initial ranking list as new queries to produce new ranking lists. The updated score of each image is calculated depending on its positions in the new ranking lists. However, in our cross-model task, the noise is not only aroused by the high dimensionality of features but also the asymmetric cross-modal neighboring relationship in the embedding space. We believe that $k$-reciprocal nearest neighbor is superior to $k$-nn on re-ranking, because it considers both issues to eliminate the influence of noises. $k$-reciprocal nearest neighbor is first defined by [38] for object retrieval and has been extend to person re-identification by [44]. In this manuscript, we first analysis and prove the asymmetric cross-modal neighboring relationship between the visual and semantic aspects of the embedding space, and then extend the $k$-reciprocal nearest neighbors to multi-modal space. As a result, the retrieved results are more related to the probe in bi-directions of semantic and visual aspects, which effectively improves the accuracy of retrieval.

III. CROSS-MODAL RETRIEVAL BOOSTED BY MVSE++ AND RECIPROCAL NEIGHBORS

In this section, we first propose a multiple feature based visual-semantic embedding method, based on which the initial cross-modal retrieval is conducted. Then, we re-rank the initial results based on MVSE++ with $k$-reciprocal nearest neighbor method. The whole framework is illustrated as Fig.1.

A. MVSE++

In original VSE++, the visual feature is extracted by CNN based object classifiers trained on ImageNet dataset [45]. Thus, the features mainly focus on objects in images. However, images and their corresponding captions usually involve the scene context besides objects. Thus, the embedding space of VSE++ is not proper for the visual semantic alignment between images and captions. Fortunately, Zhou et al. [46] selected 365 scene categories with more than 4,000 images in each category to create Places365-standard dataset, which offers an ecosystem of visual context to guide the progress on scene understanding. Therefore, we incorporate scene features extracted by CNN based scene classifiers trained on Places365 dataset. We use class activation mapping (CAM) [47] to visualize the extracted object-related and scene-related regions in Fig.2, which illustrates the complementarity of those features. It is obvious that the captions include the information of both the object and scene-context. Thus, we concatenate them to construct the multiple visual

FIGURE 2. The first row gives the original images. The second row shows the corresponding CAMs trained on ImageNet. The third row shows the corresponding CAMs trained on Places365. The fourth row shows the corresponding captions, where the words in red correspond to objects and the blue colored words correspond to scene context.
where $\phi_{obj}$ is the object-related feature and $\phi_{pla}$ is the scene-related feature. $\theta$ represents the parameters in CNNs, which extract the visual features. $\text{Norm}(\cdot)$ is a normalization operation, which guarantees the features lie on a unit hypersphere in $\phi_{multi}$.

To achieve a better alignment between visual feature and semantic feature, we map the concatenated visual feature $\phi_{multi}$ into the joint embedding space (i.e. MVSE++ space), and get the embedded multiple visual feature $f_{multi}$, as follows

$$f_{multi}(I; \phi_{obj}, \phi_{pla}) = \text{Norm}(\phi_{obj}(I; \theta_{obj})), \text{Norm}(\phi_{pla}(I; \theta_{pla})).$$

where $\phi_{obj}$ is the object-related feature and $\phi_{pla}$ is the scene-related feature. $\theta$ represents the parameters in CNNs, which extract the visual features. $\text{Norm}(\cdot)$ is a normalization operation, which guarantees the features lie on a unit hypersphere in $\phi_{multi}$.

In the training process, given the image and caption pairs $(I, C)$, we define a similarity measure function $s_{multi}(I, C)$, which aims to score the positive pairs higher than the negative pairs (making $s$ equivalent to cosine similarity).

$$s_{multi}(I, C) = f_{multi}(I; \theta_{obj}) \cdot g(C; W_{g}, \theta_{g})$$

We denote $I'$ and $C'$ as the hardest negative image and caption samples, which are the most similar negative sample to its corresponding probe. The sum of violation loss function used in [16] is actually minimizing the mean of the non-negative terms, while, in fact, the hardest negative sample is more important to R@K. Thus, our final loss function of a single example pair is denoted as follows

$$\ell(I, C) = \max_{C} \left[ \alpha + s_{multi}(I, C') - s_{multi}(I, C) \right]_{+} + \max_{I} \left[ \alpha + s_{multi}(I', C) - s_{multi}(I, C) \right]_{+},$$

where $\alpha$ represents the margin, and $(I, C)$ is the positive pair. $(I', C)$ and $(I, C')$ are the negative pairs. $[x]_{+}$ is a shorthand notation of $[x]_{+} \equiv \max(x, 0)$.

We use SGD to learn our parameters $\theta$. We take the maximum loss over each mini-batch. Thus, $e(\theta, s)$ defined as Eq.(6) is the empirical loss of all samples, which is parametrized by $\theta$.

$$e(\theta, s) = \frac{1}{N} \sum_{n=1}^{N} \ell(I_{n}, C_{n})$$

In the above process, we complement the scene context feature to the object feature in visual representation of $I$. It provides a more comprehensive representation of the visual content of objects and scene context in images. Thus, we achieve better alignment to $C'$.

## B. THE CROSS-MODAL $k$-RECIPROCAL NEAREST NEIGHBOR BASED RE-RANKING

Captioning is to describe the visual information of images with highly concentrated natural language. Captions usually describe salient objects and their related context in images. Thus, the information maintained in the visual feature and semantic feature is asymmetric. We build an embedding space of MVSE++ to align them better than that of VSE++.

In the embedding space, the similarity $s_{multi}(I, C)$ is the dot product between $I$ and $C$. Thus, it is symmetric, viz. $s_{multi}(I, C) = s_{multi}(C, I)$. However, the neighboring relationships are asymmetric. To explain clearly, we define the cross-modal $k$-nearest neighbors (i.e. the top-$k$ list) of an image vector $I$ as $\text{top}_k(I, I)$, which includes the $k$-most similar caption vectors $\{C_1, C_2, \cdots, C_k\}$ in the embedding space of MVSE++. Similarly, $I \in \text{top}_k(C, C)$ means $I$ is the top-$k$ cross-modal neighbor of $C$ in the embedding space. The asymmetric neighboring relationship means that $I \in \text{top}_k(C, C)$ does not imply $I \in \text{top}_k(I, I)$. The cross-modal retrieval is bi-directional, the nearest neighboring relationship searching from visual or semantic side in the embedding space to another modal is asymmetric, which is illustrated as Fig.3. The top-2 nearest neighbors of $I_1$ are $C_1$ and $C_2$. $I_1$ belongs to the top-2 nearest neighbors of $C_1$. However, $I_1$ does not belong to top-2 nearest neighbors of $C_2$. The traditional $k$-nn retrieval only searches from one aspect, which may suffer from this asymmetric issue. Intuitively, we deal with this issue with reciprocal retrieval.

We are not the first to use the reciprocal nearest neighbors to solve the asymmetric nearest neighbor relationship. Zhong et al. used re-ranking to improve the accuracy of person ReID [44]. Qin et al. use $k$-reciprocal nearest neighbors to retrieve object [38]. Yet in the cross-modal retrieval community, limited effort has been devoted to solve the asymmetric nearest neighbor relationship in the embedding latent space. This $k$-reciprocal nearest neighbors can be extended.
to most cross-modal retrieval. An underlying assumption of this $k$-reciprocal nearest neighbors based re-ranking is that if a returned image/caption ranks within the $k$-nearest neighbors of the probe, it is likely to be a true match, which can be used for the subsequent re-ranking.

We re-rank the initial ranking list based on the $k$-reciprocal nearest neighbors in MVSE++ space to reduce the influence of asymmetric issue. We define the cross-modal reciprocal nearest neighbors $R(I, k_1, k_2)$ of the probe image $I$ as follows

$$R(I, k_1, k_2) = \{ C_i | (C_i \in \text{top}(k_1, I)) \land (I \in \text{top}(k_2, C_i)) \}$$

(7)

where $C_i, i \in \{1, \cdots, k\}$ are the $k_1$-nearest captions of $I$ in the embedding space. To acquire the cross-modal reciprocal neighbors, we restrict $I$ belongs to the $k_2$-nearest images to caption $C_j$. The cross-modal reciprocal nearest neighbors $R(C, k_1, k_2)$ of caption $C$ can be defined similarly. Fig. 4 gives an example, for the probe image $I$, we search its top-$k_1$ neighboring captions $\{C_1, C_2, \cdots, C_{k_1}\}$ in the embedding space of MVSE++. Then, for each caption $C_j$ we retrieve its top-$k_2$ neighboring images in the embedding space of MVSE++. If the retrieved images include image $I$, we say $C_j$ is the reciprocal neighbor of $I$.

The cross-modal $k$-reciprocal nearest neighborhood relationship $k_1, k_2, I$ is a much stronger indicator of similarity than the undirection nearest neighborhood relationship $\text{top}(k_1, I)$, because it takes into account the local densities of vectors around $I$ and $C$. Consequently, the asymmetric of nearest neighborhood relationships in the cross-modal retrieval is reduced. Therefore, the re-ranking based on cross-modal reciprocal neighbor can effectively improve the performance of retrieval. We denote our retrieval method based the reciprocal re-ranking as R-MVSE++.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASET AND EVALUATION METRIC

To verify the performance of our method, we conduct experiments on Flickr30K dataset [48] and Microsoft COCO dataset [49]. Flickr30K includes 30,000 images. Following Karpathy and Fei-Fei [50], we use 1,000 images for validation, 1,000 images for testing, and the rest images for training. There are 164,062 images in MSCOCO. We also use 1,000 images for validation and 1,000 images for testing. For fair comparison, we follow Faghri et al. [17] to split the training, validation and test sets. We use 30,504 images left in the validation set of MSCOCO together with the original images for training (113,287 training images in total) to further improve the accuracy. We refer to this set as $rV$. Each image in these two datasets is annotated with 5 sentences by Amazon Mechanical Turk (AMT). We evaluate the retrieval performance using Recall@K (Viz. R@K), which is the mean number of corrections within the top-$k$ retrieved results. We also report the median rank of the closest ground truth results from the ranked list (i.e. Med r).

B. TRAINING

To compare with existing methods, we also conduct experiments on three popular visual feature encoders: VGG19 [51], ResNet152 [52] and DenseNet161 [53]. The image features are extracted by first resizing the image to $256 \times 256$. Then, the image is cropped to $224 \times 224$ using a random crop. We extract the image features by pre-computing the FC7 features (the penultimate fully connected layer) of VGG19, pre-computing the Pool5 features (the last pooling layer) of ResNet152 and pre-computing the Pool5 features (the last pooling layer) of DenseNet161. The dimensionality of image
embedding for each encoder is 4,096 for VGG19, 2,048 for ResNet152 and 2,208 for DenseNet161. We use a GRU as caption encoder, and set the dimensionality of the word vector as 300. We set the dimensionality of GRU and the joint embedding space as 1,024.

### C. ABSTRACTION STUDY

We conduct two groups of experiments on ablation study. One is to compare the performance of embedding spaces of VSE++ and MVSE++. The other is to evaluate the influence of \( k \) value on R@1 and R@10.

#### 1) VERIFICATION OF THE DIFFERENT FEATURES ON EMBEDDING

The verification of the effectiveness of our multiple features on retrieval on the validation set of MSCOCO is presented in Table 1. As shown in Table 1, both VSE++ and MVSE++ perform better on ResNet152 and DenseNet161 compared with those on VGG19, which proves the powerfulness of ResNet and DenseNet on visual feature extraction. The results of ResNet152 and DenseNet161 are comparable for both VSE++ and MVSE++. MVSE++ obtains 2.8% improvement in R@1 for caption retrieval and 2.7% improvement in R@1 for image retrieval. Thus, the results imply that our multiple features effectively improve the accuracy of the cross-modal retrieval. The performance of MVSE++ on ResNet152 and DenseNet161 are comparable, which are better than that of our method on VGG19, which is consistent with the results on MSCOCO. The results prove that the scene-related feature achieves better alignment between visual and semantic modals in the embedding space. The visualization in Fig.2 of the focus of object-related and scene-related visual features by classic CAM [47] also proves the complementary of them. Thus, the concatenated feature provides a comprehensive representation of image.

#### 2) VERIFICATION OF THE NUMBER OF NEIGHBORS

We evaluate the influence of \( k_2 \) on R@1 and R@10 accuracies on VGG19, DenseNet161 and ResNet152, which is illustrated in Fig.IV-C.2. In our experiments, we find that when \( k_1 = 30 \), the retrieval achieves the best results, so we fix \( k_1 \) at 30 in the following experiments. We can see that for caption retrieval, R@1 achieves the best performance when \( k_2 = 1 \). R@1 accuracy is stable with the increase of \( k_2 \). R@10 achieves the best performance with \( k_2 \) equals to 2 or 3, which means the probe caption usually exists in the top-3 reciprocal neighbors of the cross-modal retrieval images. For image retrieval, when \( k_2 = 2 \), R@1 achieves the best results. However, the accuracy of R@10 continues to rise until \( k_2 \) reaches to 8 in Fig.IV-C.2. Since captions are highly condensed semantic information of images, images contain more information than captions. Once caption is given, the corresponding image is easy to be determined. However, image can be described by more captions. Thus, the image description is objective. That is why dense caption [63] and caption variation [64] are popular recently.

### TABLE 1. The effect of different features on embedding on MSCOCO.

| Model                   | Caption Retrieval | Image Retrieval |
|-------------------------|-------------------|-----------------|
|                         | R@1   | R@10  | Med r | R@1   | R@10  | Med r |
| VSE++(VGG19)            | 45.9  | 88.1  | 2.0   | 36.8  | 82.1  | 2.0   |
| VSE++(DenseNet161)      | 56.0  | 93.0  | 1.0   | 41.4  | 85.3  | 2.0   |
| VSE++(ResNet152)        | 58.5  | 93.3  | 1.0   | 42.9  | 86.8  | 2.0   |
| MVSE++(VGG19)           | 49.3  | 89.4  | 2.0   | 39.2  | 83.3  | 2.0   |
| MVSE++(DenseNet161)     | 59.9  | 94.4  | 1.0   | 45.7  | 89.9  | 2.0   |
| MVSE++(ResNet152)       | 61.9  | 94.9  | 1.0   | 46.3  | 89.1  | 2.0   |

### TABLE 2. The effect of different features on embedding on Flickr30K.

| Model                   | Caption Retrieval | Image Retrieval |
|-------------------------|-------------------|-----------------|
|                         | R@1   | R@10  | Med r | R@1   | R@10  | Med r |
| VSE++(VGG19)            | 38.6  | 74.6  | 2.0   | 26.8  | 66.8  | 4.0   |
| VSE++(DenseNet161)      | 44.3  | 83.2  | 2.0   | 31.6  | 73.9  | 3.0   |
| VSE++(ResNet152)        | 43.7  | 82.1  | 2.0   | 32.3  | 72.1  | 3.0   |
| MVSE++(VGG19)           | 39.2  | 79.0  | 2.0   | 29.0  | 70.3  | 4.0   |
| MVSE++(DenseNet161)     | 47.1  | 85.2  | 2.0   | 34.3  | 75.8  | 3.0   |
| MVSE++(ResNet152)       | 50.4  | 84.5  | 1.0   | 36.7  | 76.0  | 3.0   |

### TABLE 3. The effect of re-ranking to MVSE++ on MSCOCO. We fix \( k_1 \) at 30, and \( k_2 \) at 2.

| Model                   | Caption Retrieval | Image Retrieval |
|-------------------------|-------------------|-----------------|
|                         | R@1   | R@10  | Med r | R@1   | R@10  | Med r |
| MVSE++(VGG19)           | 49.3  | 89.4  | 2.0   | 39.2  | 83.3  | 2.0   |
| MVSE++(DenseNet161)     | 59.9  | 94.4  | 1.0   | 45.7  | 89.9  | 2.0   |
| MVSE++(ResNet152)       | 61.9  | 94.9  | 1.0   | 46.3  | 89.1  | 2.0   |
| R-MVSE++ (VGG19)        | 50.8  | 90.4  | 1.0   | 53    | 86.8  | 1.0   |
| R-MVSE++(DenseNet161)   | 60.7  | 95.1  | 1.0   | 62.3  | 90.5  | 1.0   |
| R-MVSE++(ResNet152)     | 62.6  | 96.1  | 1.0   | 63.4  | 92.3  | 1.0   |

### TABLE 4. The effect of re-ranking to MVSE++ on Flickr30K. We fix \( k_1 \) at 30, and \( k_2 \) at 2.

| Model                   | Caption Retrieval | Image Retrieval |
|-------------------------|-------------------|-----------------|
|                         | R@1   | R@10  | Med r | R@1   | R@10  | Med r |
| MVSE++(VGG19)           | 39.2  | 79.0  | 2.0   | 29.0  | 70.3  | 4.0   |
| MVSE++(DenseNet161)     | 47.1  | 85.2  | 2.0   | 34.3  | 75.8  | 3.0   |
| MVSE++(ResNet152)       | 50.4  | 84.5  | 1.0   | 36.7  | 76.0  | 3.0   |
| R-MVSE++ (VGG19)        | 40.1  | 80.8  | 2.0   | 38.8  | 75.1  | 2.0   |
| R-MVSE++(DenseNet161)   | 50.8  | 86.3  | 1.0   | 48.8  | 80.9  | 2.0   |
| R-MVSE++(ResNet152)     | 53.8  | 86.4  | 1.0   | 50.7  | 81.8  | 1.0   |
TABLE 5. Results of experiments on MSCOCO.

| Model                  | Trainset    | Caption Retrieval | Image Retrieval |
|------------------------|-------------|-------------------|-----------------|
|                         |             | R@1  | R@10 | Med r  | R@1  | R@10 | Med r  |
| Order [54]             | rV          | 46.7 | 88.9 | 2.0    | 37.9 | 85.9 | 2.0    |
| Embedding Network [18] | ?           | 50.4 | 69.4 | -      | 39.8 | 86.6 | -      |
| sm-LSTM [33]           | ?           | 53.2 | 91.5 | 1.0    | 40.7 | 87.4 | 2.0    |
| 2WayNet [55]           | ?           | 55.8 | -    | -      | 39.7 | -    | -      |
| GMM-HGLLM [56]         | ?           | 55.9 | 93.3 | 2      | 44.0 | 94.6 | 2      |
| SPE [14]               | ?           | 50.1 | 89.2 | -      | 39.6 | 86.9 | -      |
| GXN [35]               | rV          | 68.5 | 97.9 | -      | 56.7 | 94.5 | -      |
| SCO [57]               | rV          | 69.9 | 97.5 | -      | 56.7 | 94.8 | -      |
| SCAN [58]              | rV          | 72.7 | 98.4 | -      | 58.8 | 94.8 | -      |
| CMPM(ResNet-152) [59]  | ?           | 56.1 | 92.9 | -      | 44.6 | 89.0 | -      |
| JGCAr [60]             | ?           | 52.7 | 90.5 | -      | 40.2 | 85.7 | -      |
| VSRN [61]              | ?           | 76.2 | 98.2 | -      | 62.8 | 95.1 | -      |
| MTFN [62]              | ?           | 74.3 | 97.9 | -      | 60.1 | 95.0 | -      |
| VSE++(ResNet152) [17]  | rV          | 58.5 | 93.3 | 1.0    | 42.9 | 86.8 | 2.0    |
| R-MVSE++(ResNet152)    | rV          | 62.6 | 96.1 | 1.0    | 63.4 | 92.3 | 1.0    |
| VSE++(ResNet152,fine-tuned) [17] | rV | 64.6 | 95.7 | 1.0    | 52.0 | 92.0 | 1.0    |
| R-MVSE++(DenseNet161, fine-tuned) | rV | 69.6 | 96.8 | 1.0    | **66.9** | **95.5** | **1.0** |

According to above results, we fix $k_1$ at 30 and $k_2$ at 2, and then compare MVSE++ and R-MVSE++ to demonstrate the effectiveness of reciprocal re-ranking. In Tab.3, we can see that our re-ranking method achieves evident improvement for retrieval. Especially for image retrieval, our method achieves improvements at 13.8% in R@1 on VGG19, 17.1% on ResNet152, 16.6% on DenseNet161 and 14.4% on fine-tuned DenseNet161. Tab.4 illustrates the results of re-ranking on Flickr30 with $k_1$ at 30 and $k_2$ at 2. We see that our method achieves improvement at 4.8% in R@1 on VGG19, 5.8% on ResNet152, and 5.1% on DenseNet161 for caption retrieval. The improvement on image retrieval is higher than that of caption retrieval. Scene-related feature of image helps us find the better aligned captions, which in turn helps us find the images similar to the probe on both object and scene. However, once the scene-semantic is missing in the probe caption, it is not easy to recover it through re-ranking.

D. VALIDATION ON RETRIEVAL

In previous experiments, we only update the parameters of captions $\theta_\psi$ and the mapping metrics $W_f, W_g$. Considering the encoders of visual feature are pre-trained for classification task, in order to make the encoders better suit our visual-semantic embedding task, we fine-tune partial parameters. We report the fine-tuned results in Tab.5 to compare our embedding method with the methods Order [54], Embedding Network [18], sm-LSTM [33], 2WayNet [55], GMM-HGLLM [56] and SPE [14]. Our method achieves best results over those methods. We see that our method on fine-tuned DenseNet161 achieves 5.0% improvement in R@1 for caption retrieval and 14.9% improvement on image retrieval over that of VSE++ on fine-tuned ResNet152, which means DenseNet161 can be better fine-tuned to obtain the more reasonable parameters for visual feature encoding, towarding cross-modal embedding. Fig.IV-C.2 illustrates the superiority of our method over VSE++. We also compare our method with these methods in Tab.6 on Flickr30K dataset. Our embedding method on DenseNet161 with fine-tuning achieves 59.6% accuracy in R@1 for image retrieval and 56.1% accuracy in R@1 for caption retrieval, which outperforms these methods.

Comparing with the more recent state-of-the-art approaches: GXN [35], SCAN [58], SCO [57], CMPM(ResNet-152) [59], JGCAr [60], VSRN [61] and MTFN [62]. The results in Tab.5 and Tab.6 indicate that our method achieve comparative performance with these
Table 6. Results of experiments on Flickr30K.

| Model                        | Caption Retrieval | Image Retrieval |
|------------------------------|-------------------|-----------------|
|                             | R@1   | R@10  | Med r | R@1   | R@10  | Med r |
| m-CNN [65]                  | 33.6  | 74.9  | 3.0   | 26.2  | 69.6  | 4.0   |
| Embedding Network [18]       | 40.7  | 79.2  | -     | 29.2  | 71.7  | -     |
| DCCA [10]                   | 27.9  | 68.2  | 4.0   | 26.8  | 66.9  | 4.0   |
| sem-LSTM [33]               | 42.5  | 81.9  | 2.0   | 30.2  | 72.3  | 3.0   |
| GMH-HLM [56]                | 36.5  | 73.3  | -     | 24.7  | 66.8  | 7.0   |
| SPE [14]                    | 40.3  | 79.9  | -     | 29.7  | 72.1  | -     |
| DAN(VGG19) [34]             | 41.4  | 82.5  | 2.0   | 31.8  | 72.5  | 3.0   |
| DAN(ResNet152) [34]         | 55.0  | 89.0  | 1.0   | 39.4  | 79.1  | 3.0   |
| SCo [57]                    | 55.5  | 89.3  | -     | 41.1  | 80.1  | -     |
| SCAN [58]                   | 63.6  | 93.7  | -     | 43.2  | 82.3  | -     |
| CPM(ResNet152) [59]         | 48.3  | 64.5  | -     | 35.7  | 74.1  | -     |
| JGCar [60]                  | 44.9  | 82.7  | -     | 35.2  | 72.4  | -     |
| VSRN [61]                   | **71.3** | **96.0** | - | **54.7** | **88.2** | - |
| MTFN [62]                   | 65.3  | 93.3  | -     | 52.0  | 86.1  | -     |
| VSE++(VGG19,fine-tuned) [17] | 41.3  | 77.9  | 2.0   | 31.4  | 71.2  | 3.0   |
| R-MVSE++(VGG19,fine-tuned)  | 48.1  | 83.5  | 2.0   | 45.4  | 79.3  | 2.0   |
| VSE++(DenseNet161,fine-tuned) [17] | 51.2  | 86.3  | 1.0   | 38.4  | 78.0  | 2.0   |
| R-MVSE++(DenseNet161,fine-tuned) [17] | 59.6  | 89.1  | 1.0   | 56.1  | 83.8  | 1.0   |
| VSE++(ResNet152,fine-tuned) | 52.9  | 87.2  | 1.0   | 39.6  | 79.5  | 2.0   |
| R-MVSE++(ResNet152,fine-tuned) [17] | 62.5  | 90.4  | 1.0   | **58.9** | **86.5** | **1.0** |

**FIGURE 6.** Examples of testing images and the top-1 retrieved captions by our model and VSE++ on ResNet152. The value in brackets is the rank of the highest ranked ground-truth caption.

V. CONCLUSION

In this paper, we investigate the image-caption retrieval, which joints the object-scene feature based visual-semantic embedding and reciprocal neighbors. We show that the multi-feature, which avoids the visual-semantic miss-alignment on both objects and scene context aspects, is more representative than the previously used single feature for image-caption retrieval. We also propose to incorporate the re-ranking method based on k-reciprocal nearest neighbors to re-rank the initial rank list, which symmetrizes nearest neighborhood relationships in the cross-modal retrieval. We evaluate our method on MSCOCO and Flickr30K datasets. The results consistently prove the superiority of our method over the state-of-the-art methods. Besides retrieval, our visual-semantic embedding model can be applied to image captioning and VQA tasks, etc. Considering failure retrieval cases of the proposed method are always caused by the interaction between different objects, we will focus on how to well caputre the local alignment between visual and semantic information in the future.

REFERENCES

[1] V. Ordonez, G. Kulkami, and T. L. Berg, “Im2Text: Describing images using 1 million captioned photographs,” in Proc. Adv. Neural Inf. Process. Syst., Meeting Held, Granada, Spain, Dec. 2011, pp. 1143–1151.

[2] R. Mason and E. Charniak, “Nonparametric method for data-driven image captioning,” in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, vol. 2, Baltimore, MD, USA, Jun. 2014, pp. 592–598.

[3] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, “Collective generation of natural image descriptions,” in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Jeju Island, South Korea, Jul. 2012, pp. 359–368.

[4] A. Nenkova and L. Vanderwende, “The impact of frequency on summarization,” Microsoft Res., Redmond, WA, USA, Tech. Rep. MSR-TR-2005, 2005, vol. 101.

[5] L. Pang, S. Zhu, and C.-W. Ngo, “Deep multimodal learning for affective analysis and retrieval,” IEEE Trans. Multimedia, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[6] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” Neural Comput., vol. 16, no. 12, pp. 2659–2664, Dec. 2004.
Y. Cao, M. Long, J. Wang, and S. Liu, “Deep visual-semantic quantization,” Int. J. Comput. Vis., vol. 106, no. 2, pp. 210–233, Jan. 2014.

M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” J. Artif. Intell. Res., vol. 47, pp. 853–899, Aug. 2013.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in Proc. 30th Int. Conf. Mach. Learn. (ICML), Atlanta, GA, USA, Feb. 2013, pp. 1247–1255.

F. Yan and K. Mikolajczyk, “Deep correlation for matching images and text,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3441–3450.

N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep Boltzmann machines,” in Proc. Adv. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, Dec. 2012, pp. 2222–2230.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in Proc. Int. Conf. Mach. Learn., Jun. 2011, pp. 689–696.

Y. Shen, L. Liu, L. Shao, and J. Song, “Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 4117–4126.

L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 5005–5013.

X. Yang, T. Zhang, and C. Xu, “Cross-domain feature learning in multi-media,” IEEE Trans. Multimedia, vol. 17, no. 1, pp. 64–78, Jan. 2015.

R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” 2014, arXiv:1411.2539. [Online]. Available: http://arxiv.org/abs/1411.2539.

F. Faghri, D. J. Fleet, J. Ryan Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” 2017, arXiv:1707.05612. [Online]. Available: http://arxiv.org/abs/1707.05612.

L. Wang, Y. Li, J. Huang, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 394–407, Feb. 2019.

T. Gupta, K. Shih, S. Singh, and D. Hoiem, “Aligning image-word representations improves interactive transfer across vision-language tasks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 4213–4222.

J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, Jun. 2015, pp. 2625–2634.

R. Kiros, R. Salakhutdinov, and R. Zemel, “Multimodal neural language models,” in Proc. ICML, Jun. 2014, pp. 595–603.

J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, “Deep captioning with multimodal recurrent neural networks (m-RNN),” CoRR, vol. abs/1412.6632, 2015.

J. Weston, S. Bengio, and N. Usunier, “Wsabie: Scaling up to large multi-label image classification,” in Proc. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2011, pp. 2764–2772.

Y.-H.-H. Tsai, L.-K. Huang, and R. Salakhutdinov, “Learning robust visual-semantic embeddings,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3571–3580.

S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, “Identity-aware textual-visual matching with latent co-attention,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 1890–1899.

Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, Jul. 2017, pp. 3571–3580.

R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unsupervised cross-modal retrieval through adversarial learning,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2017, pp. 1–6.

H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 1063–6919.

L. He, X. Xu, H. Li, Y. Yang, F. Shen, and H. T. Shen, “Instance-aware image and sentence matching with selective multimodal LSTM,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2310–2318.

D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. van Gool, “Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors,” in Proc. CVPR, Colorado Springs, CO, USA, Jun. 2011, pp. 777–784.

X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, “Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Providence, RI, USA, Jun. 2012, pp. 3013–3020.

I. Gonzalez-Diaz, M. Birinci, F. Diaz-de-Maria, and E. J. Delp, “Neighborhood matching for image retrieval,” IEEE Trans. Multimedia, vol. 19, no. 3, pp. 544–558, Mar. 2017.

O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, “Automatic query expansion with a generative feature model for object retrieval,” in Proc. IEEE 11th Int. Conf. Comput. Vis., Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, “Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing,” IEEE Trans. Multimedia, vol. 18, no. 12, pp. 2553–2566, Dec. 2016.

R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2911–2918.

Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person identification with k-reciprocal encoding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, Jul. 2017, pp. 3652–3661.

L. Fei-Fei, “ImageNet: A large-scale hierarchical human-annotated image database,” in Proc. Conf. Comput. Vision (CVPR), Providence, RI, USA, Jun. 2009, pp. 1530–1537.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2921–2929.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Trans. Assoc. Comput. Linguistics, vol. 2, pp. 67–78, Dec. 2014.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., Zurich, Switzerland, Sep. 2014, pp. 740–755.

A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, Jun. 2015, pp. 3128–3137.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, Lille, France, Jul. 2015, pp. 1–14.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 1, no. 2, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.
[54] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, “Order-embeddings of images and language,” in Proc. ICLR, San Juan, PR, USA, May 2015, pp. 1–13.

[55] A. Eisenschat and L. Wolf, “Linking image and text with 2-way nets,” 2016, arXiv:1608.07973. [Online]. Available: http://arxiv.org/abs/1608.07973

[56] B. Klein, G. Lev, G. Sadeh, and L. Wolf, “Associating neural word embeddings with deep image representations using Fisher vectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 4437–4446.

[57] Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts and order for image and sentence matching,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6163–6171.

[58] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” 2018, arXiv:1803.08024. [Online]. Available: http://arxiv.org/abs/1803.08024

[59] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” 2018, arXiv:1803.08024. [Online]. Available: http://arxiv.org/abs/1803.08024

[60] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, “Joint global and co-attentive representation learning for image-sentence retrieval,” in Proc. ACM Multimedia Conf. Multimedia Conf. (MM), 2018.

[61] 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.

[62] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, “Matching images and text with multi-modal tensor fusion and re-ranking,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 12–20.

[63] L. Wu, Y. Wang, X. Li, and J. Gao, “Deep attention-based spatially recursive networks for fine-grained visual recognition,” IEEE Trans. Cybern., vol. 49, no. 5, pp. 1791–1802, May 2019.

[64] Y. Li, J. Shi, and D. Lin, “Low-latency video semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5997–6005.

[65] L. Ma, Z. Lu, L. Shang, and H. Li, “Multimodal convolutional neural networks for matching image and sentence,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2623–2631.

MENGMEI JIANG received the B.S. and M.S. degrees from Xidian University, Xi’an, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree. He is also with the School of Electronic and Engineering, Xidian University. His research interests include text detection, statistical machine learning, image caption, and pattern recognition.

XIANGNAN ZHANG received the B.S. and M.S. degrees from Xidian University, Xi’an, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree. He is also with the School of Electronic and Engineering, Xidian University. His research interests include text detection, statistical machine learning, image caption, and pattern recognition.

HENG LIU received the B.S. degree from Xidian University, Xi’an, China, in 2016, where he is currently pursuing the M.S. degree. He is also with the School of Electronic and Engineering, Xidian University. His research interests include image processing, statistical machine learning, image caption, and pattern recognition.

CHUNNA TIAN (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi’an, China, in 2002, 2005, and 2008, respectively. From 2006 to 2007, she was with the Visual Computing and Image Processing Laboratory, Oklahoma State University, as a Visiting Student. She is currently a Professor with the School of Electronic and Engineering, Xidian University. Her research interests include image processing and pattern recognition.

WEI WEI (Senior Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, Xi’an, China, in 2012. He is currently an Associate Professor with the School of Computer Science and Engineering, Northwestern Polytechnical University. His research interests include image processing, machine learning, and pattern recognition. He has been authored more than 40 articles, including International Journal of Computer Vision, the IEEE TRANSACTIONS ON IMAGE PROCESSING, Pattern Recognition, CVPR, ICCV, ECCV, AAAI, IJCAI, and so on.