Jointly Discriminating and Frequent Visual Representation Mining

Qiannan Wang, Ying Zhou, Zhaoyan Zhu, Xuefeng Liang (✉), and Yu Gu

School of Artificial Intelligence, Xidian University, China
xliang@xidian.edu.cn

Abstract. Discovering visual representation in an image category is a challenging issue, because the visual representation should not only be discriminating but also frequently appears in these images. Previous studies have proposed many solutions, but they all separately optimized the discrimination and frequency, which makes the solutions sub-optimal. To address this issue, we propose a method to discover the jointly discriminating and frequent visual representation, named as JDFR. To ensure discrimination, JDFR employs a classification task with cross-entropy loss. To achieve frequency, JDFR uses triplet loss to optimize within-class and between-class distance, then mines frequent visual representations in feature space. Moreover, we propose an attention module to locate the representative region in the image. Extensive experiments on four benchmark datasets (*i.e.* CIFAR10, CIFAR100-20, VOC2012-10 and Travel) show that the discovered visual representations have better discrimination and frequency than ones mined from five state-of-the-art methods with average improvements of 7.51% on accuracy and 1.88% on frequency.

1 Introduction

Visual patterns are basic visual elements that commonly appear in images and tend to convey higher-level semantics than raw pixels. Thus, mining visual patterns is a fundamental issue in computer vision, and have been applied to many vision tasks, such as object recognition [1,2], object detection [3], and scene classification [1], to name a few. The visual representation of a category is a type of visual pattern that represents the discernible regularity in the visual world and captures the essential nature of visual objects or scenes. Unlike visual pattern, visual representation does not have to appear in every image of a category. Recently, it has been utilized in the tourism industry. By mining visual representations from travel photos, users can discover useful information about tourism destinations for travel recommendation [4,5,6]. Visual representation has two properties [1,7]: 1) *discrimination*, which means it represents only a particular image category rather than the other categories; 2) *frequency*, which means it frequently appears in images of the category.

To tackle this issue, handcrafted features, *i.e.* SIFT [8] and HOG [9], were firstly used for visual pattern mining. Due to the scale invariability and tolerating

a certain distortion in a local space, they are regarded as low-level visual patterns. However, the local feature is limited in its capability of expressing the semantics of images. Recently, convolutional neural network (CNN) has been often utilized as a feature extractor [1,2,10,11], as it is able to learn the high-level semantic representation of images. The CNN features were associated with association rules [1], clustering algorithm [10], and unsupervised max-margin analysis [3] to discover visual patterns. The discrimination or frequency of these patterns was separately guaranteed through varied optimizations. In addition, other studies [12,13] applied image co-saliency detection to find the visual patterns that appear in all images.

Aforementioned methods still face two issues regarding visual representation mining. Firstly, the separation of discrimination and frequency will make the solution sub-optimal. Secondly, image co-saliency requires the representation to appear in all images.

To address the above issues, we propose a jointly discriminating and frequent visual representation mining method (JDFR) to discover the visual representations that are simultaneously discriminating and frequent. In JDFR, the end-to-end network is jointly optimized by cross-entropy loss and triplet loss. The cross-entropy loss ensures the discrimination of visual representation using a classification task. The triplet loss ensures the frequency by exploring the highly dense visual representations in the feature space, which have a close within-class distance and a sufficient between-class distance. Since the visual representation often is only a region in the image, we design an attention module to locate the most discriminating regions in the images.

Therefore, our main contributions are as follows:

- We propose an end-to-end framework jointly optimized by cross-entropy loss and triplet loss to discover the discriminating and frequent visual representations.
- We designed a channel and spatial attention module to locate the visual representations in images.
- Experiments show that our JDFR outperforms five state-of-the-arts on four benchmark datasets.

2 Related Work

2.1 Visual pattern mining

Since understanding visual pattern is a fundamental issue in visual cognition, many studies have been conducted on this issue. Handcrafted features [8,9,14,15] were first applied for visual pattern mining. Doersch C et al. [9] used the HOG descriptor to represent visual patterns, which were iteratively optimized by SVM. However, such local features can not well represent the semantic information of images well, and thus are usually regarded as low-level visual patterns.

Recently, CNN demonstrated a remarkable performance on many vision tasks [1,2,10,11], because its high-level features can represent better semantic information. Li et al. [1] extracted features from image patches using a CNN model, and

retrieved semantic patches from these features based on association rules. Since the frequency and discrimination of these patches were optimized separately, the method did not achieve a desirable performance on the classification task. Zhang et al. [10] mined visual patterns using the mean shift in a binary feature space, and thus was able to ensure the frequency. Moreover, they enhanced the discrimination by leveraging contrast images. However, their outputs were images instead of visual patterns, meanwhile, the frequency and the discrimination were optimized independently as well. Yang et al. [3] exploited the hierarchical abstraction of CNN and utilized unsupervised max-margin analysis to locate visual patterns in images. This method is effective for discrimination but cannot guarantee the frequency. An emerging study [7] is able to mine visual patterns simply by analyzing filter activations in CNN. Due to the empirical design of the methodology, there is no solid theory that supports the two aforementioned properties of visual representation, especially the frequency. Furthermore, other studies [16,17,18] have shown that the frequently occurring images could be found by clustering methods.

One can see that none of aforementioned methods is able to jointly optimize the discrimination and the frequency. By contrast, we propose a new framework (JDFR) that is able to mine the best discriminating and frequent representations using the joint optimization.

2.2 Image co-saliency detection

Some studies [19,20,21,22] consider the visual pattern mining as an image co-saliency problem, which refers to detecting the common salient objects or regions in a set of relevant images. Image co-saliency detection methods can be grouped into three categories: bottom-up, fusion-based and learning-based methods. Bottom-up methods score image regions based on feature priors to simulate visual attention. Fu et al. [23] proposed three visual attention cues including contrast, spatial and corresponding ones. Later, they proposed a two-stage propagation framework using background and foreground cues [19]. Fusion-based methods ensembled the detection results of existing saliency or co-saliency methods. For example, Cao et al. [24] obtained the self-adaptive weight via a rank constraint to combine the co-saliency maps. Huang et al. [25] used multiscale superpixels to jointly detect salient object via low-rank analysis. Studies [26,27] have discovered inter-image correspondence through the high-level semantic features extracted from CNNs. Learning based methods have developed significantly in recent years because in the breakthrough of deep learning models [20,21,28]. Wei et al. [29] proposed an end-to-end framework based on the Masked-guided FCN to discover co-salient objects. Ge et al. [28] proposed an unsupervised CNN to jointly optimize the co-saliency maps.

However, there is a significant difference between image co-saliency detection and visual representation mining. Image co-saliency requires the same pattern appearing in all images. Instead, visual representation is a pattern that represents the major characteristic of the category, not necessarily to appear in each image.

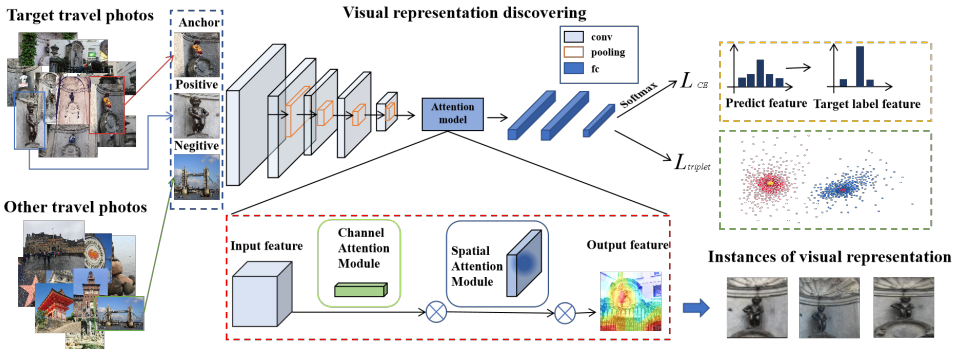

Fig. 1: The schematic diagram of our JDFR. Best viewed in color.

3 The Proposed Method

Since visual representations should be the most discriminating and frequent image regions in an image category, we firstly employ a classification task to discover the discriminating images. For frequency, we then apply the triplet loss to make features of the visual representation of one category close, but features belonging to different categories far in the feature space. Finally, we utilize an attention module to locate the representative region in each image. The schematic diagram of our method is illustrated in Fig.1.

3.1 The classification task for discriminating images

The images are discriminating for a category when they represent the characteristics of the data in the category well. Therefore, discovering discriminating images can be considered as a classification task. The more discriminating the images are, the higher classification accuracy they achieve. We define a classification network with parameters as $f(\cdot)$. Given an image x and its label y, the network predicts its label $\hat{y} = f(x)$ that indicates which category x belongs to. Please refer to the yellow dashed box in Fig. 1. To optimize the network, we employ the cross-entropy loss $L_{CE}(y, \hat{y})$ that is

$$L_{CE} = \sum_{j=1}^{m} \sum_{i=1}^{n} -y_{ji} log(\hat{y}_{ji}) - (1 - y_{ji}) log(1 - \hat{y}_{ji}), \tag{1}$$

where m is the number of categories, n is the number of images in a category, y_{ji} denotes the ground truth, and $\hat{y}_{ji}$ denotes the output of the network.

3.2 The triplet loss for frequent visual representation

The frequent visual representation of a category should represent the majority of the data and its features should have a highly dense distribution. In other

words, the features of frequent visual representation should have smaller distance between each other. The cosine similarity function is commonly used to measure the similarity between features [30]. By setting an appropriate threshold for cosine loss, the feature distance learned by the network can be less than the threshold. However, cosine similarity only guarantees the within-class distance, which is inappropriate for our task. Because the visual representation may not always appear in every image of the category, the distances among the discriminating features of the category should not be constrained by a fixed and small threshold. Instead, mining visual representation requires not only a proper within-class distance, but also a sufficient between-class distance, as shown in the green dashed box in Fig. 1. To this end, we use triplet loss to ensure that an image x_i^a(anchor) in the i-th category is closer to other samples x_i^p (positive) than the image x_j^n (negative) in the j-th category [31], as shown in the blue dashed box in Fig. 1. This can be formulated as:

$$\|g(x_i^a) - g(x_i^p)\|_2^2 + \alpha < \|g(x_i^a) - g(x_j^n)\|_2^2 \qquad i \neq j, \tag{2}$$

where $g(x)$ denotes the feature vector of the sample x, and α is the margin (between-class distance) between positive and negative pairs. In this work, $g(x)$ is the high-level feature in the network because it involves the semantic information. We use Euclidean distance to measure the similarity between the features. The network is optimized by

$$L_{triplet} = \sum_{i=1}^{t} [\|g(x_i^a) - g(x_i^p)\|_2^2 - \|g(x_i^a) - g(x_j^n)\|_2^2 + \alpha]_+, \tag{3}$$

where, t denotes the number of triples. By using triplet loss, the mined visual representations can be frequent.

3.3 Attention modules for locating visual representation

The visual representation, is often a region in the image rather than the whole image, especially for the image that contains multiple objects. Meanwhile, it is expected to be discriminating for the category. To locate the most discriminating region in an image, we must address two problems: 1) How to find the discriminating object in each image? 2) How to locate the most discriminating region?

For the first problem, since different channels in the feature map F focuses on different objects in the image, we design a channel attention module that explores the inter-channel relationship of features to find the most discriminating object in each image. Average-pooling, $AvgPool_s$, is applied on the intermediate feature map, F, to aggregate spatial information. Max-pooling, $MaxPool_s$, is applied on F to aggregate distinctive object features. In order to obtain a finer channel-wise attention, we put them into a multi-layer perceptron (MLP) with one hidden layer to produce a channel attention map $M_c \in \mathbb{R}^{(c \times 1 \times 1)}$, which

shows the weight of each channel. It can be formulated as:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool_s(F)) + MLP(MaxPool_s(F))) \\ &= \sigma(W_1(W_0(F_c^{avg})) + W_1(W_0(F_c^{max}))), \end{aligned} \tag{4}$$

$$F' = M_c(F) \otimes F, \tag{5}$$

where σ is the sigmoid function, $W_0 \in \mathbb{R}^{(c/r \times c)}$ and $W_1 \in \mathbb{R}^{(c \times c/r)}$ are the MLP weights, r is the reduction ratio, $F_c^{avg} \in \mathbb{R}^{(c \times 1 \times 1)}$ and $F_c^{max} \in \mathbb{R}^{(c \times 1 \times 1)}$ denote the average-pooled features and the max-pooled features, respectively, $\otimes$ denotes the element-wise multiplication, and F' is the channel feature map.

For the second problem, we design a spatial attention module that explores the inter-spatial relationship of the features. Due to the effectiveness of pooling operations along the channel axis in highlighting informative regions [32], we apply average pooling, $AvgPool_c$, and max-pooling, $MaxPool_c$, along the channel axis to locate the most discriminating region on F'. They are then concatenated and convolved by a standard convolution filter. Please refer to the red dashed box in Fig. 1. It can be formulated as:

$$M_s(F') = \sigma(f^{7\times7}([AvgPool_c(F'); MaxPool_c(F')])), \tag{6}$$

$$F'' = M_s(F') \otimes F', \tag{7}$$

where $f^{7\times7}$ is a convolution operation with the size of 7×7, and F'' is the feature map with attention.

To form the attention module, the channel attention and spatial attention are combined in a sequential manner with channel first order. We place it following the last convolution layer in the network. After training, the optimized attention module can locate the most discriminating region in each image, as shown in Fig.2.

3.4 The unified model and optimization

To ensure both discrimination and frequency of visual representations, we jointly optimize the network using the cross-entropy loss and triplet loss. The overall objective function is

$$\min_{\theta} \mathrm{L} = \beta L_{triplet} + \gamma L_{CE}, \tag{8}$$

where β and γ are constants to balance the contributions of the two losses.

4 EXPERIMENTS

In this section, we firstly describe the experiment set-up, including datasets, implementation details, and evaluation metrics. Then, we examine the effectiveness of our method by comparing it with five state-of-the-arts on four benchmark datasets. Finally, we conduct ablation studies by controlling major influence factors.

4.1 Experiment Set-up

Datasets. Four datasets are selected in our experiment. First, we choose three benchmark datasets that are commonly used for visual pattern mining evaluation, there are CIFAR-10 [33], CIFAR-100-20 [33], and VOC2012-10 [34]. Since many categories in CIFAR100 are very challenging for visual representation mining, we select 20 categories from it, named as CIFAR100-20. VOC2012 is originally designed for object detection, so all images contain multiple objects. In many cases, the shared objects are too small to be representative. Therefore, we select 10 categories with representative objects from VOC2012, named as VOC2012-10. The travel photo dataset is collected from the popular travel website, TripAdvisor[1]. Photos from one travel destination belong to a category. Details of datasets are shown in Table 1. The Test sets are used for discovering the visual representation, and the Train and Validation sets with category labels are used for training our model.

Table 1: Details of four datasets.

dataset	CIFAR10	CIFAR100-20	VOC2012-10	Travel
Category	10	20	10	20
Train	40000	8000	4905	64904
Validation	10000	2000	786	16227
Test	10000	2000	948	20000

Implementation Details. In this work, all the experiments were implemented by PyTorch on an NVIDIA 2080Ti with 11GB of on-board memory. We fine-tuned the pre-trained VGG-19 on each training set. The VGG-19 was jointly trained by cross-entropy loss and triplet loss, and was optimized using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.1. To balance the two loss functions, the hyperparameters β and γ were set to 0.3 and 0.7 respectively. The training stopped when no significant reduction of the validation error occurred, about 50 epochs. To find the most frequent representations, we applied a density-based clustering algorithm for mining task, and the number of features with the highest density was set as $N_u = 20$. Thus, 20 visual instances would be discovered for the visual representation from each category.

Competing Methods. The recent study [1] reported that the CNN-based visual patterns mining methods had largely outperformed the traditional hand-crafted based methods. Therefore, we compared JDFR with five state-of-the-arts for performance evaluation, which are all CNN-based methods. They are (1) Mid-level Deep Pattern Mining (MDPM) [1], (2)Contrastive Binary Mean Shift (CBMS) [10], (3) Part-level Convolutional Neural Network model (P-CNN) [3], (4) PatternNet [7] and (5) Masked-guided FCN (MFCN) [35]. Since only the code of MDPM was available from authors, we strictly implemented other methods

[1] https://www.tripadvisor.com

according to their papers. The CNNs in P-CNN, PatternNet and MFCN were fine-tuned on the training set as well. P-CNN was trained with cross-entropy loss and SGD with momentum technique and the initial learning rate of 1e-3 that was the best rate in our tests. Since the detailed setting of experimental parameters were not given in the PatternNet, we trained it with MSE loss and Adam with the best initial learning rate of 1e-4, while the experimental parameters of MFCN used were consistent with the original paper.

4.2 Evaluation Metrics

The following two metrics were employed to evaluate the effectiveness of our model:

Discrimination Evaluation. Previous works for visual pattern mining used the image classification task as a proxy to evaluate their performances. Thus, we followed their protocol and trained a Resnet50 for classifying images to the corresponding categories, which is for evaluating the discrimination of the discovered visual representations. The results were an average accuracy and F1-score of those 20 instances of visual representation retrieved by the clustering step. Since MDPM divided the input image to a set of patches for subsequent processing, it was evaluated on the retrieved visual patches.

Frequency Evaluation. Intuitively, the discrimination cannot evaluate the frequency of discovered visual representation directly. Few previous studies explicitly measured it either. In this paper, we proposed a new metric (Frequency rate, FR) to compute the percentage of the images that are similar to the discovered visual representations in the high-level feature space. This is defined as:

$$\mathrm{FR} = \frac{1}{N_w \times N_u \times N} \sum_{w=1}^{N_w} \sum_{u=1}^{N_u} \sum_{v=1}^{N} \left\| S_{u,v}^{w} \geq T_f \right\|_0 , \tag{9}$$

where, $S_{u,v}^{w} = cos(p_u^w, p_v^w)$ is a cosine similarity, p_u^w and p_v^w are the feature maps coming from the last convolution layer of aforementioned ResNet50. p_u^w is the feature map of one image from $w\text{-}th$ category, and p_v^w is the feature map of an instance of discovered visual representation from $w\text{-}th$ category. N_w, N_u, and N are the number of categories, the number of images in each category, and the number of retrieved instances of visual representation(s), respectively. T_f denotes the similarity threshold. In this work, it was set three levels: 0.866, 0.906, and 0.940, which are corresponding to 30°, 25°, and 20° between two feature vectors, respectively.

4.3 Result and Analysis

Quantitative results and comparison For the discrimination of visual representation, classification accuracy and F1-score are used for evaluation. The results of JDFR and five state-of-the-arts on four datasets are listed in Table 2. It shows that JDFR outperforms other five competing methods. MDPM performs the worst because it divides the image into patches, which could lose some

Table 2: Comparison among six approaches on discrimination of discovered visual representations.

Method	CIFAR10		CIRAR100-20		VOC2012-10		Travel	
	mAcc	F1	mAcc	F1	mAcc	F1	mAcc	F1
MDPM[1]	0.7820	0.7800	0.7800	0.7750	0.8000	0.8100	0.7800	0.8530
CBMS[10]	0.8800	0.8790	0.8630	0.8640	0.8650	0.8663	0.8630	0.9550
P-CNN[3]	0.8850	0.8880	0.8800	0.8750	0.8720	0.8715	0.8800	0.9725
PatternNet[7]	0.8300	0.8330	0.8200	0.8250	0.7950	0.7963	0.8200	0.9374
MFCN[35]	0.8450	0.8440	0.8434	0.8452	0.8312	0.8416	0.8434	0.9057
JDFR(Ours)	**0.9650**	**0.9650**	**0.9450**	**0.9440**	**0.9100**	**0.9100**	**0.9975**	**0.9975**

Table 3: Comparison of six approaches on frequency of discovered representations at three thresholds T_f.

Datasets	T_f	MDPM[1]	CBMS[10]	P-CNN[3]	PatternNet[7]	MFCN[35]	JDFR(Ours)
CIFAR10	0.940(20°)	0.1025	0.1565	0.1533	0.1473	0.1516	**0.1733**
	0.906(25°)	0.3289	0.5169	0.5224	0.4856	0.5003	**0.5429**
	0.866(30°)	0.5673	0.8689	0.8792	0.8386	0.8515	**0.9122**
CIFAR100-20	0.940(20°)	0.0322	0.0425	0.0489	0.0406	0.0372	**0.0523**
	0.906(25°)	0.1540	0.2039	0.2153	0.2002	0.2037	**0.2346**
	0.866(30°)	0.3370	0.5511	0.5314	0.5468	0.5531	**0.5880**
VOC2012-10	0.940(20°)	0.0425	0.0665	0.0725	0.0627	0.0780	**0.0923**
	0.906(25°)	0.1489	0.2410	0.2312	0.2245	**0.2468**	0.2415
	0.866(30°)	0.3556	0.4789	0.4456	0.4289	0.4774	**0.4876**
Travel	0.940(20°)	0.0752	0.1277	0.1000	0.1163	0.1325	**0.1668**
	0.906(25°)	0.1428	0.2603	0.3002	0.2375	0.2470	**0.3348**
	0.866(30°)	0.2555	0.4109	0.4355	0.3926	0.4058	**0.5076**

semantic information. Surprisingly, PatternNet concentrates on mining discriminating patterns, but achieves the second worse performance, the reason might be that the discriminating information of their result is only provided by one max-pooling layer (last convolution), which lacks of adequate high-level semantic features. MFCN reaches the third worse since it is designed for co-saliency detection, which requires the visual pattern must appear in all images. P-CNN achieves the second best because P-CNN is more robust than other methods by using multi-scale information of images. JDFR performs the best on all datasets due to optimizing both discrimination and frequency of the visual representations. Compared with P-CNN, JDFR only improves 3.8% accuracy and 3.85% F1-score on VOC2012-10, but improves 8.0% accuracy and 7.7% F1-score on CIFAR10. Moreover, one can see that most methods perform better on VOC2012-10 and Travel than CIFAR10 and CIFAR100-20. The possible explanation is that our network has a fixed architecture of Conv module, and the feature map in the high-level layer becomes rather small when the input is small. Thus, the resolution of images in CIFAR10 and CIFAR100-20 is much smaller than the one in VOC2012-10 and Travel, which makes high-level features of images in CIFAR contain less semantics. In addition, all methods perform better on CI-

FAR10 than CIFAR100-20. We can observe that the number of training images in CIFAR100-20 is smaller than CIFAR10, which can make the trained network sub-optimal.

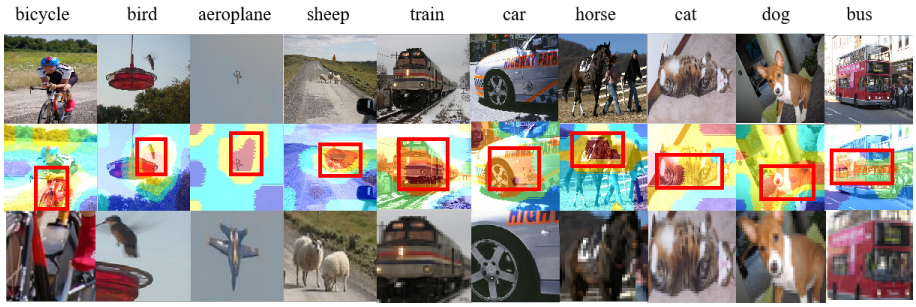

Fig. 2: Instances of discovered visual representations by JDFR from VOC2012-10. The first row lists the original images, the second row shows the attention maps after joint optimization, and the last row demonstrates the discovered instances of visual representation. Best viewed in color.

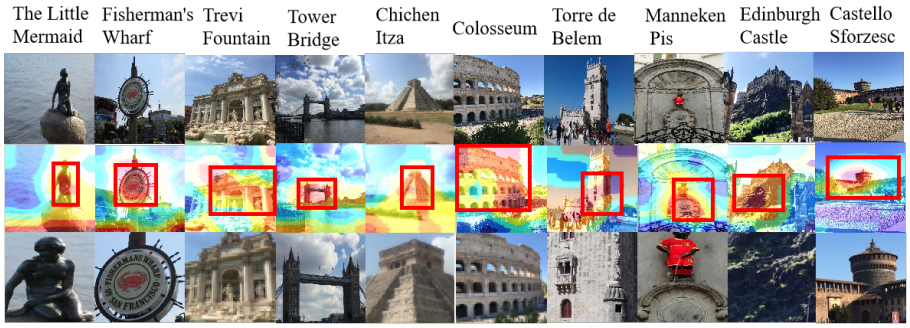

Fig. 3: Instances of discovered visual representations by JDFR from Travel. Best viewed in color.

Secondly, we compare the frequency of discovered visual representations at a varied threshold T_f on four datasets. The result is similar to the classifcation evaluation, as shown in Table 3. The visual representations discovered by JDFR are more frequent than ones from other methods at almost all thresholds. CBMS discovers the most frequent representation on VOC2012-10 when $T_f = 30°$. But it is just slightly higher than ours. Although MDPM uses a frequent pattern mining algorithm, it still performs the worst. PatternNet and P-CNN focus on mining the discriminating patterns, while their results also have high frequencies. All

the above results demonstrate that our JDFR can discover better discriminating and frequent visual representations.

Qualitative results and comparison To subjectively evaluate the performance of our method, we illustrate the attention maps and discovered visual representations of ten categories in VOC2012-10 in Fig.2 and ten tourism destinations in Travel Fig.3, respectively. One can see that they contain the symbolic content and represent these categories well.

Fig. 4: Instances of visual representation of aeroplane category in VOC2012-10 discovered by (a) MDPM [1], (b) CBMS [10], (c) P-CNN [3], (d) PatternNet [7], (e) MFCN [35] and (f) JDFR (ours), respectively. Best viewed in color.

For the qualitative comparison, we list ten instances of visual representation discovered by six approaches from aeroplane category in VOC2012-10, as shown in Fig.4, and Manneken Pis in Travel, as shown in Fig.5, respectively. One can observe that MDPM produces the worst result marked with the blue box, because it utilizes image patches that may merely have a part of the symbolic object. CBMS finds the frequent images in the yellow box instead of the visual representation in the images. P-CNN and PatternNet are able to discover the visual representations but include a few off-target errors only marked with the red boxes. MFCN is designed to mine the co-existing objects across all images. Thus, it finds the same object in dataset highlighted in green box, but does not work on images which include other objects only. By contrast, our method can retrieve consistent instances of visual representations.

Fig. 5: Instances of visual representation of Manneken Pis in Travel photo dataset discovered by (a) MDPM [1], (b) CBMS [10], (c) P-CNN [3], (d) PatternNet [7], (e) MFCN [35] and (f) JDFR (ours), respectively. Best viewed in color.

4.4 Ablative study

Variants of model To further verify our main contributions, we firstly compare JDFR with the model only with cross-entropy loss and only with triplet loss, respectively. Here, we set $\alpha = 1.5$ and $T_f = 20^\circ$ because 20° is the strictest threshold of similarity measure. Results are listed in Table 4.

Table 4: Ablation study of JDFR on four datasets.

	mAcc			F1			FR 0.940(20°)		
	CE	Triplet	CE+Triplet	CE	Triplet	CE+Triplet	CE	Triplet	CE+Triplet
CIFAR10	0.9100	0.8850	**0.9650**	0.9100	0.8850	**0.9650**	0.1667	0.1731	**0.1893**
CIFAR100-20	0.9050	0.8500	**0.9450**	0.9050	0.8490	**0.9440**	0.0405	0.0365	**0.0482**
VOC2012-10	**0.9100**	0.8350	**0.9100**	**0.9100**	0.8270	**0.9100**	0.0650	0.0634	**0.0715**
Travel	0.9800	0.9650	**0.9975**	0.9800	0.9640	**0.9975**	0.1464	0.1569	**0.1668**

One can observe JDFR achieves the best frequent and discriminating performance on four datasets. Specifically, JDFR improves 1.6% frequency on CIFAR10 compared with the Triplet model. The average improvement on all datasets is 1.15%. Analogously, JDFR improves 5.5% accuracy and 5.5% F1-score on CIFAR10 compared with the CE model. On average, JDFR raise the accuracy 2.8% and F1-score 2.8%. These results demonstrate that joint optimization does improve the both frequency and discrimination of visual representations.

(a) car (b) bicycle (c) bus (d) dog

Fig. 6: Four images from VOC2012-10 with labels of car, bicycle, bus and dog, respectively. The labels are indicated by yellow box. The shared object is shown in red box.

Moreover, we find the frequency on VOC2012-10 increases the least among four datasets, and the accuracy is not improved. This might be caused by the characteristics of VOC2012-10. Each image in VOC2012-10 has multiple objects which may be shared with images belonging to other categories. For instance, Fig. 6(a) and Fig. 6(b) are labeled by car and bicycle, respectively. But they both include car and bicycle. Figure 6(c) and Fig. 6(d) are other examples that contain both bus and dog but have different labels. This suggests that many images, which are in different categories, may have very similar semantic features. In this case, triplet loss can not contribute much for the classification task. Thus, the discrimination performance of JDFR on VOC2012-10 is identical to the one of CE model. Since it does enlarge the between-class distance, the performance of frequency is still improved.

Table 5: Comparison of JDFR performance with different α.

	mAcc			F1			FR 0.940(20°)		
	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$
CIFAR10	0.9600	**0.9650**	0.9600	0.9600	**0.9650**	0.9590	0.1934	0.1899	**0.2042**
CIFAR100-20	0.9280	**0.9450**	0.9400	0.9270	**0.9440**	0.9380	**0.0556**	0.0482	0.0508
VOC2012-10	0.8950	0.9100	**0.9200**	0.8960	0.9100	**0.9150**	0.0691	0.0715	**0.0742**
Travel	**0.9975**	**0.9975**	0.9925	**0.9975**	**0.9975**	0.9925	**0.1843**	0.1668	0.1657

Varied margin α The margin α can adjust the between-class distance. We test varied α and show the results in Table 5. Due to the characteristics of VOC2012-10, the larger value of α can improve both frequency and discrimination of visual representations. However, an overlarge α is not appropriate to other data. One can observe that the best overall performance on all datasets can be reached when $\alpha = 1.5$. Therefore, We choose this setting for all experiments in this paper.

5 Conclusion

In this work, we propose a jointly discriminating and frequent visual representation mining method (JDFR) to address the problem of discovering visual representations. Unlike previous studies focusing on either the discriminating patterns or frequent patterns, JDFR can optimize both the discrimination and frequency of discovered visual representations simultaneously. Moreover, our channel and spatial attention modules help to locate the representations in images. To evaluate the effectiveness of JDFR, we conduct experiments on four diverse datasets. The results of classification accuracy and frequency demonstrate that JDFR is able to discover the best visual representation in comparing with five state-of-the-art methods.

6 Acknowledgments

This work is supported by the Science and Technology Plan of Xi'an (20191122015KYPT011JC013), the Fundamental Research Funds of the Central Universities of China (No. JX18001) and the Science Basis Research Program in Shaanxi Province of China (No. 2020JQ-321, 2019JQ-663).

References

1. Li, Y., Liu, L., Shen, C., Van Den Hengel, A.: Mining mid-level visual patterns with deep cnn activations. IJCV **121** (2017) 344–364
2. Chen, Z., Maffra, F., Sa, I., Chli, M.: Only look once, mining distinctive landmarks from convnet for visual place recognition. In: IROS. (2017) 9–16
3. Yang, L., Xie, X., Lai, J.: Learning discriminative visual elements using part-based convolutional neural network. Neurocomputing **316** (2018) 135–143
4. Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I., Chen, G.: Travel recommendation using geo-tagged photos in social media for tourist. Wireless Personal Communications **80** (2015) 1347–1362
5. Vu, H.Q., Li, G., Law, R., Ye, B.H.: Exploring the travel behaviors of inbound tourists to hong kong using geotagged photos. Tourism Management **46** (2015) 222–232
6. Bronner, F., De Hoog, R.: Vacationers and ewom: Who posts, and why, where, and what? Journal of Travel Research **50** (2011) 15–26
7. Li, H., Ellis, J.G., Zhang, L., Chang, S.F.: Automatic visual pattern mining from categorical image dataset. International Journal of Multimedia Information Retrieval **8** (2019) 35–45
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV. Volume 2. (1999) 1150–1157
9. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? Communications of the ACM **58** (2015) 103–110
10. Zhang, W., Cao, X., Wang, R., Guo, Y., Chen, Z.: Binarized mode seeking for scalable visual pattern discovery. In: CVPR. (2017) 3864–3872
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)

12. Tan, Z., Liang, W., Wei, F., Pun, C.M.: Image co-saliency detection by propagating superpixel affinities. In: ICASSP. (2013)
13. Chang, K.Y., Liu, T.L., Lai, S.H.: From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In: CVPR. (2011) 2129–2136
14. Zhang, B., Gao, Y., Zhao, S., Liu, J.: Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. IEEE TIP **19** (2009) 533–544
15. Kim, S., Jin, X., Han, J.: Disiclass: discriminative frequent pattern-based image classification. In: KDD Workshop on Multimedia Data Mining. (2010)
16. Lapuschkin, S., Binder, A., Montavon, G., Muller, K.R., Samek, W.: Analyzing classifiers: Fisher vectors and deep neural networks. In: CVPR. (2016) 2912–2920
17. Gong, Y., Pawlowski, M., Yang, F., Brandy, L., Bourdev, L., Fergus, R.: Web scale photo hash clustering on a single machine. In: CVPR. (2015) 19–27
18. Chum, O., Matas, J.: Large-scale discovery of spatially related images. IEEE TPAMI **32** (2009) 371–377
19. Fu, H., Xu, D., Lin, S., Liu, J.: Object-based rgbd image co-segmentation with mutex constraint. In: CVPR. (2015) 4428–4436
20. Fu, H., Xu, D., Zhang, B., Lin, S.: Object-based multiple foreground video co-segmentation. In: CVPR. (2014) 3166–3173
21. Tang, K., Joulin, A., Li, L.J., Fei-Fei, L.: Co-localization in real-world images. In: CVPR. (2014) 1464–1471
22. Wei, L., Zhao, S., Bourahla, O.E.F., Li, X., Wu, F.: Group-wise deep co-saliency detection. arXiv:1707.07381 (2017)
23. Fu, H., Cao, X., Tu, Z.: Cluster-based co-saliency detection. IEEE TIP **22** (2013) 3766–3778
24. Cao, X., Tao, Z., Zhang, B., Fu, H., Feng, W.: Self-adaptively weighted co-saliency detection via rank constraint. IEEE TIP **23** (2014) 4175–4186
25. Han, J., Cheng, G., Li, Z., Zhang, D.: A unified metric learning-based framework for co-saliency detection. IEEE TCSVT **28** (2017) 2473–2483
26. Luo, Y., Jiang, M., Wong, Y., Zhao, Q.: Multi-camera saliency. IEEE TPAMI **37** (2015) 2057–2070
27. Bors, A.G., Papushoy, A.: Image retrieval based on query by saliency content. In: Visual Content Indexing and Retrieval with Psycho-Visual Models. Springer (2017) 171–209
28. Ge, C., Fu, K., Liu, F., Bai, L., Yang, J.: Co-saliency detection via inter and intra saliency propagation. Signal Processing: Image Communication **44** (2016) 69–83
29. Li, H., Ngan, K.N.: A co-saliency model of image pairs. IEEE TIP **20** (2011) 3365–3375
30. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: ACCV. (2010) 709–720
31. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. (2015) 815–823
32. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:1612.03928 (2016)
33. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. (2009)
34. Shetty, S.: Application of convolutional neural network for image classification on pascal voc challenge 2012 dataset. arXiv:1607.03785 (2016)

35. Zhang, K., Li, T., Liu, B., Liu, Q.: Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In: CVPR. (2019) 3095–3104