2019 IEEE/CVF International Conference on Computer Vision (ICCV)

Transferable Contrastive Network for Generalized Zero-Shot Learning

Huajie Jiang[1,2,3,4], Ruiping Wang[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Shanghai Institute of Microsystem and Information Technology, CAS, Shanghai, 200050, China
[4]School of Information Science and Technology, ShanghaiTech University, Shanghai, 200031, China
huajie.jiang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Abstract

Zero-shot learning (ZSL) is a challenging problem that aims to recognize the target categories without seen data, where semantic information is leveraged to transfer knowledge from some source classes. Although ZSL has made great progress in recent years, most existing approaches are easy to overfit the sources classes in generalized zero-shot learning (GZSL) task, which indicates that they learn little knowledge about target classes. To tackle such problem, we propose a novel Transferable Contrastive Network (TCN) that explicitly transfers knowledge from the source classes to the target classes. It automatically contrasts one image with different classes to judge whether they are consistent or not. By exploiting the class similarities to make knowledge transfer from source images to similar target classes, our approach is more robust to recognize the target images. Experiments on five benchmark datasets show the superiority of our approach for GZSL.

1. Introduction

Object recognition is one of the basic issues in computer vision. It has made great progress in recent years with the rapid development of deep learning approaches [18, 33, 30, 13], where large numbers of labeled images are required, such as ImageNet [28]. However, collecting and annotating large numbers of images are difficult, especially for fine-grained categories in specific domains. Moreover, such supervised learning approaches can only recognize a fixed number of categories, which is not flexible. In contrast, humans can learn from only a few samples or even recognize unseen objects. Therefore, learning visual classifiers with no need of human annotation is becoming a hot topic in recent years.

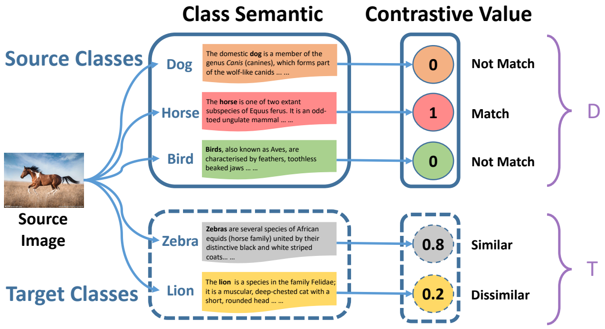
Class Semantic
Contrastive Value
Source Classes
Dog
The domestic **dog** is a member of the genus *Canis* (canines), which forms part of the wolf-like canids ... ...
0
Not Match
Horse
The **horse** is one of two extant subspecies of Equus ferus. It is an odd-toed ungulate mammal ... ...
1
Match
Bird
Birds, also known as Aves, are characterised by feathers, toothless beaked jaws ... ...
0
Not Match
D
Source Image
Zebra
Zebras are several species of African equids (horse family) united by their distinctive black and white striped coats... ...
0.8
Similar
Target Classes
Lion
The **lion** is a species in the family Felidae; it is a muscular, deep-chested cat with a short, rounded head ... ...
0.2
Dissimilar
T

Figure 1. Illustration diagram that shows the motivations of transferable contrastive learning. The training images should not only match their class semantics (discriminative property) but also have relatively high contrastive values with similar target classes (transfer property). 'D' represents discriminative learning and 'T' represents transfer learning.

Zero-shot learning (ZSL) aims to learn classifiers for the target categories where no labeled images are accessible. It is accomplished by transferring knowledge from the source categories with the help of semantic information. Semantic information can build up the relations among different classes thus to enable knowledge transfer from source classes to target classes. Currently the most widely used semantic information includes attributes [19, 9] and word vectors [10, 2]. Traditional ZSL approaches usually learn universal visual-semantic transformations among the source classes and then apply them to the target classes. In this way, the visual samples and class semantics can be projected into a common space, where zero-shot recognition is conducted by the nearest neighbor approach.

Although ZSL has made great progress in recent years, the strong assumption that the test images only come from the target classes is not realistic for practical applications. Therefore, generalized zero-shot learning (GZSL) [6, 39] draws much attention recently, where test samples may come from either source or target classes. However, most

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00986

existing ZSL approaches perform badly on GZSL task because they are easy to overfit the source classes, which indicates that they learn little knowledge about the target classes. These approaches learn the models only on the source categories and ignore the targets. Since the domain shift problem exists [11], the models learned on the source classes may not be suitable to the target classes, which results in overfitting the source categories in the GZSL task.

In order to tackle such problem, we propose to explicitly transfer the knowledge from the source classes to the target categories. The key problem for ZSL is that no labeled images are available for the target categories so we could not directly train the target image classifiers. An intuitive idea is to learn target classifiers from similar source images. For example, we could leverage the source images 'horse' to learn the target class 'zebra'. Based on this idea, we propose a novel transferable contrastive network for generalized zero-shot learning. It automatically contrasts the images with class semantics to judge whether they are consistent or not. Figure 1 shows the motivations of our approach, where two key properties for ZSL are considered in the contrastive learning process: discriminative property and transferable property. We maximize the contrastive values of images with corresponding class semantics and minimize the inconsistent ones among source classes thus to ensure that our model is discriminative enough to recognize different classes. Furthermore, to make the contrast transferable to the target classes, we utilize the class similarities to transfer knowledge from the source-class images to similar target classes. In this way, the model will be more robust to the target categories though no labeled target images are available to learn the model.

The main contributions of this paper are in two aspects. First, we propose a novel transferable contrastive network for GZSL, where a new network structure is designed for contrastive learning. Second, we consider both the discriminative property and transferable property in the contrastive learning procedure, where the discriminative property ensures to effectively discriminate different classes and the transferable property guarantees the robustness to the target classes. Experiments on five benchmark datasets show the superiority of the proposed approach.

2. Related Work

2.1. Semantic Information

Semantic information is the key to ZSL. It builds up the relations between the source and target classes thus to enable knowledge transfer. Recently, the most widely used semantic information in ZSL is attributes [19, 9] and word vectors [22]. Attributes are general descriptions of objects. They are accurate but need human experts for definition and annotation. Word vectors are automatically learned from large numbers of text corpus which reduces human labor. However, there is much noise in the texts, which restricts their performance. In this paper, we use the attributes as the semantic information since they are more accurate to bridge the source and target classes.

2.2. Visual-Semantic Transformations

Visual-semantic transformations establish relationships between the visual space and the semantic space. According to different projection directions, current ZSL approaches can be grouped into three types: visual to semantic embeddings, semantic to visual embeddings, latent space embeddings. We will introduce them in detail below.

Visual to semantic embeddings. These approaches learn the transformations from the visual space to the semantic space and perform image recognition in the semantic space. In the early age of ZSL, [19, 9] propose to learn attribute classifiers to transfer knowledge from the source to the target classes. They train each attribute classifier independently, which is time-consuming. To tackle such problem, [1, 2] consider all attributes as a whole and learn label embedding functions to maximize the compatibilities between images and corresponding class semantics. Furthermore, [24] proposes to synthesize the semantic representations of test images by a convex combination of source-class semantics using the probability outputs of source classifiers. To learn more robust transformations, [23] proposes a deep neural network to combine attribute classifier learning and semantic label embedding.

Semantic to visual embeddings. These approaches learn the transformations from semantic space to the visual space and perform image recognition in the visual space, which can effectively tackle the hubness problem in ZSL [8, 29]. [5, 21, 40] predict the visual samplers by learning embedding functions from the semantic space to the visual space. [27] adds some regularizers to learn the embedding function from class semantic to corresponding visual classifiers and [35] utilizes knowledge graphs to learn the same embedding functions. Some other works directly synthesize the target-class classifiers [4] or learn the target-class prototypes [14] in the visual space by utilizing the class structure information. [17, 7] exploit the auto-encoder framework to learn both the semantic to visual and visual to semantic embeddings simultaneously. Inspired by the generative adversarial networks, [38] generates the target-class samples in the feature space and directly learns the target classifiers.

Latent space embedding. These approaches encode the visual space and semantic space into a latent space for more effective image recognition. Since the predefined semantic information may be not discriminative enough to classify different classes, [41, 42] propose to use class similarities as the embedding space and [16] proposes discriminative latent attributes for zero-shot recognition. Moreover, [3] exploits metric learning techniques, where relative distance is utilized, to improve the embedding models. In order to learn robust visual-semantic transformations, [10, 31, 26, 23] utilize deep neural networks to project the visual space and the semantic space into a common latent space and align the representations of the same class.

Our approach belongs to the latent space embedding, but there is a little difference. Traditional methods aim to minimize the distance of images and corresponding class semantics in the latent space for image recognition, while our approach fuses their information for contrastive learning.

2.3. Zero-Shot Recognition

Zero-shot recognition is the last step for ZSL, most of which can be grouped into two categories. The distance-based approaches usually exploit the nearest neighbour approach to recognize the target-class samples [1, 12, 41, 16] and the classifier-based approaches directly learn the visual classifiers to recognize the target-class images [4, 38]. Our approach utilizes contrastive values for image recognition.

2.4. Discussions about Relevant Works

Most existing approaches ignore the target classes when learning the recognition model, so they are prone to overfitting the source classes in GZSL task. To tackle this problem, [38, 43] leverage the semantic information of target classes to generate image features for training target classifiers. Although satisfactory performance has been achieved, it is difficult to train and use the generative models. While our approach is easy to learn. Moreover, it is complementary to such generative approaches. [20] proposes a calibration network that calibrates the confidence of source classes and uncertainty of target classes. Different from it, we directly transfer knowledge to the target classes, which is more effective for GZSL. [11] uses all the unlabeled target images to adjust the models in transductive ZSL settings. However, these images are often unavailable in practical conditions, so we perform the inductive ZSL task. [15] proposes adaptive metric learning to make the model suitable for the target classes. However, the linear model restricts its performance. Another relevant work is [32], which also studies the relations between images and class semantics. Compared with [32], we design a novel network structure for TCN. Moreover, we explicitly transfer knowledge from the source images to similar target classes, which makes our model more robust to the target categories.

3. Approach

The objective of our approach is learning how to contrast the image with the class semantics. Figure 2 shows the general framework of the proposed transferable contrastive network (TCN). It contains two parts: information fusion and contrastive learning. Instead of computing the distance between images and class semantics using fixed metric for recognition, we fuse their information and learn a metric that automatically judges whether the fusions are consistent or not, where high contrastive values should be obtained between images and corresponding class semantics. In order to make the contrastive mechanism suitable to the target classes, we explicitly transfer knowledge from source images to similar target classes since no target images are available for training. More details will be described below.

3.1. Problem Settings

In zero-shot learning, we are given K source classes (denoted as $\mathcal{Y}^s$) and L target classes (denoted as $\mathcal{Y}^t$), where the source and target classes are disjoint, *i.e.* $\mathcal{Y}^s \cap \mathcal{Y}^t = \emptyset$. We use the index $\{1, ..., K\}$ to represent the source classes and $\{K+1, ..., K+L\}$ to represent the target classes. The source classes contain N labeled images $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}^s\}_{i=1}^{N}$, while no labeled images are available for the target classes. $\mathcal{X}$ represents the visual sample space. To build up the relations between the source and target classes, semantic information $\mathcal{A} = \{\boldsymbol{a}_c\}_{c=1}^{K+L}$ is provided for each class $c \in \mathcal{Y}^s \cup \mathcal{Y}^t$. The goal of ZSL is to learn visual classifiers of target classes $f_{zsl} : \mathcal{X} \rightarrow \mathcal{Y}^t$ and the goal of GZSL is to learn more general visual classifiers of all classes $f_{gzsl} : \mathcal{X} \rightarrow \mathcal{Y}^s \cup \mathcal{Y}^t$.

3.2. Contrastive Network

Information Fusion. An intuitive way of contrasting an image with one class is to fuse their information and judge how consistent the fusion is. Therefore, we first encode the images and class semantics into the same latent feature space to fuse their information. As is shown in Figure 2, we use two branches of neural network to encode the image and the class semantics into the same feature space respectively, where convolutional neural network (CNN) is utilized to encode the images and the multilayer perceptrons (MLP) is utilized to encode the class semantic information (attributes or word vectors). Then an element-wise product operation ($\otimes$) is exploited to fuse the information from these two domains. Let $f(\boldsymbol{x}_i)$ denote the coding feature of the ith image and $g(\boldsymbol{a}_j)$ represent the coding feature of the jth class semantic, we can get the fused feature $\boldsymbol{z}_{ij}$ as:

$$\boldsymbol{z}_{ij} = f(\boldsymbol{x}_i) \otimes g(\boldsymbol{a}_j) \tag{1}$$

where $\boldsymbol{a}_j$ is the class semantic of the jth class. Then we can feed $\boldsymbol{z}_{ij}$ to the next stage to judge how well the image i is consistent with class j.

Contrastive Learning. Different from previous approaches that use fixed distance, such as Euclidean distance or cosine distance, to compute the similarities between images and classes for image recognition, we design a contrastive network that automatically judges how well the image is consistent with a specific class. Let v_{ij} denote the

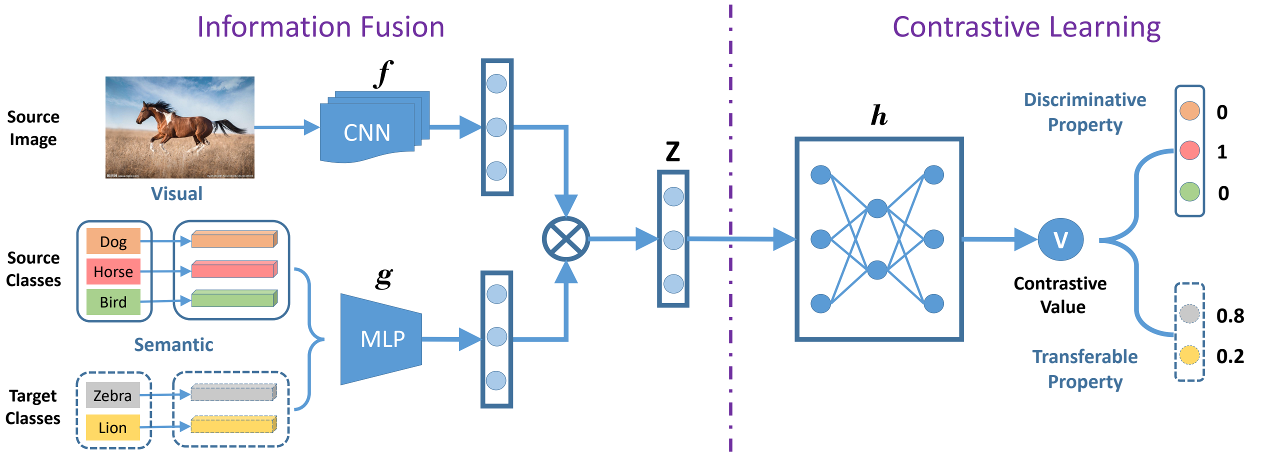

Figure 2. The framework of transferable contrastive network. The information fusion module merges the image information with the class semantic information. The contrastive learning module automatically judges whether the fusion is consistent or not. '$\otimes$' denotes the element-wise product operation.

contrastive value between image i and class j, we can obtain it from the fused feature $\boldsymbol{z}_{ij}$ as:

$$v_{ij} = h(\boldsymbol{z}_{ij}) \tag{2}$$

where h is the contrastive learning function.

In the contrastive learning phase, we should consider two characters: discriminative property and transferable property. Discriminative property indicates that the contrastive model should be discriminative enough to classify different classes. Transferable property means that the contrastive model should be generalized to the target classes.

In order to enable the discriminative property, we utilize the semantic information of source classes as supervision, where the contrastive values of consistent fusions are maximized and those of inconsistent ones are minimized. The loss function can be formulated by the cross-entropy loss:

$$\mathcal{L}_D = -\sum_{i=1}^{N}\sum_{j=1}^{K} m_{ij}\log v_{ij} + (1-m_{ij})\log(1-v_{ij}) \tag{3}$$

where m_{ij} is a class indicator. Let y_i be the class label for the ith image, then m_{ij} can be obtained by:

$$m_{ij} = \begin{cases} 1, & y_i = j \\ 0, & y_i \neq j \end{cases} \tag{4}$$

The goal of ZSL is to recognize the target classes. If we only use the source classes in the contrastive learning phase, it is easy to overfit and the model would be less transferable to the target classes. This is the problem that exists in most ZSL approaches. Unfortunately, we don't have labeled target images to take part in the contrastive learning process. To tackle such problem, we explicitly transfer knowledge from source images to the target classes by class similarities. In other words, the source images could also be utilized to learn similar target classes. Let s_{kj} denote the similarity of source class k (k=1,...,K) to target class j (j=K+1,...,K+L) and then the loss function for transferable property is formulated as:

$$\mathcal{L}_T = -\sum_{i=1}^{N}\sum_{j=K+1}^{K+L} s_{y_ij}\log v_{ij} + (1-s_{y_ij})\log(1-v_{ij}) \tag{5}$$

To summarize, our full loss function is:

$$\mathcal{L} = \mathcal{L}_D + \alpha\mathcal{L}_T \tag{6}$$

where α is a parameter that controls the relative importance of discriminative property and transferable property.

3.3. Class Similarity

In order to accomplish the contrastive learning approach proposed above, the similarities between the source and target classes should be obtained. Inspired by the sparse coding approach, we utilize the target classes to reconstruct a source class and the reconstruction coefficients are viewed as the similarity of the source class to the target classes. The objective function is:

$$\boldsymbol{s}_k = \arg\min_{\boldsymbol{s}_k} ||\boldsymbol{a}_k - \sum_{j=K+1}^{K+L}\boldsymbol{a}_j s_{kj}||_2^2 + \beta||\boldsymbol{s}_k||_2 \tag{7}$$

where $\boldsymbol{a}_k$ is the semantic information of class k and s_{kj} is the jth element of $\boldsymbol{s}_k$, which denotes the similarities of source class k to target class j. β is the regularization parameter. Then we normalize the similarity by

$$s_{kj} = \frac{s_{kj}}{\sum_{j=K+1}^{K+L} s_{kj}} \tag{8}$$

Dataset	*Img*	*Attr*	*Source*	*Target*
APY [9]	15,339	64	15 + 5	12
AWA1 [19]	30,475	85	27 + 13	10
AWA2 [37]	37,322	85	27 + 13	10
CUB [34]	11,788	312	100 + 50	50
SUN [25]	14,340	102	580 + 65	72

Table 1. Statistics for attribute datasets: APY , AWA1, AWA2, CUB and SUN in terms of image numbers (*Img*), attribute numbers (*Attr*), training + validation source class numbers (*Source*) and target class numbers (*Target*).

3.4. Zero-Shot Recognition

We conduct zero-shot recognition by comparing the contrastive values of one image with all the class semantics.

For ZSL, we classify one image to the class which has the largest contrastive value among target classes, which can be formulated as:

$$P_{zsl}(x_i) = \max_j \{v_{ij}\}_{j=K+1}^{K+L} \tag{9}$$

For GZSL, we classify one image to the class which has the largest contrastive value among all classes, which can be formulated as:

$$P_{gzsl}(x_i) = \max_j \{v_{ij}\}_{j=1}^{K+L} \tag{10}$$

4. Experiment

4.1. Datasets and Settings

We conduct experiments on five widely used ZSL datasets: APY [9], AWA (2 versions AWA1 [19] and AWA2 [37]), CUB [34], SUN [25]. APY is a small-scale coarse-grained dataset with 64 attributes, which contains 20 object classes of aPascal and 12 object classes of aYahoo. AWA1 is a medium-scale animal dataset which contains 50 animal classes with 85 attributes annotated. AWA2 is collected by [37], which has the same classes as AWA1. CUB is a fine-grained and medium-scale dataset, which contains 200 different types of birds annotated with 312 attributes. SUN is a medium-scale dataset containing 717 types of scenes where 102 attributes are annotated. In order to make fair comparisons with other approaches, we conduct our experiment on the more reasonable pure ZSL settings recently proposed by [37]. The details of each dataset and class splits for source and target classes are shown in Table 1.

Implementation Details. We extract the image features $f(\boldsymbol{x})$ by the ResNet101 model [13] and use class attributes as the semantic information. The class semantic transformation $g(\boldsymbol{a})$ is implemented by a two-layer fully connected neural network, where the hidden layer dimension is set to 1024 and the output size is 2048. The contrastive learning $h(\boldsymbol{z})$ is also implemented by the fully connected neural network, where the hidden dimension is 1024 and the output size is 1. We use Leaky ReLU as the nonlinear activation function for all the hidden layers and sigmoid function for the last layer [1]. The hyperparameter α is fine-tuned in the range [0.001, 0.01, 0.1, 1] by the validation set.

Method	**APY**	**AWA1**	**AWA2**	**CUB**	**SUN**
DAP [19]	33.8	44.1	46.1	40.0	39.9
IAP [19]	36.6	35.9	35.9	24.0	19.4
CONSE [24]	26.9	45.6	44.5	34.3	38.8
CMT [31]	28.0	39.5	37.9	34.6	39.9
SSE [41]	34.0	60.1	61.0	43.9	51.5
LATEM [36]	35.2	55.1	55.8	49.3	55.3
ALE [1]	39.7	59.9	62.5	54.9	58.1
DEVISE [10]	39.8	54.2	59.7	52.0	56.5
SJE [2]	32.9	65.6	61.9	53.9	53.7
EZSL [27]	38.3	58.2	58.6	53.9	54.5
SYNC [4]	23.9	54.0	46.6	55.6	56.3
SAE [17]	8.3	53.0	54.1	33.3	40.3
CDL [14]	43.0	69.9	-	54.5	**63.6**
RNet [32]	-	68.2	64.2	55.6	-
FGN [38]	-	68.2	-	57.3	60.8
GAZSL [43]	41.1	68.2	70.2	55.8	61.3
DCN [20]	**43.6**	65.2	-	56.2	61.8
TCN (ours)	38.9	**70.3**	**71.2**	**59.5**	61.5

Table 2. Zero-shot recognition results on APY, AWA1, AWA2, CUB and SUN (%). '-' denotes that the results are not reported.

4.2. Performance on ZSL and GZSL

To demonstrate the effectiveness of the transferable contrastive network, we compare our approach with several state-of-the-art approaches. Table 2 shows the comparison results of ZSL, where the performance is evaluated by the average per-class top-1 accuracy. It can be seen that our approach achieves the best performance on three datasets and is comparable to the best approach on SUN, which indicates that transferable contrastive network can make good knowledge transfer to the target classes. Our approach is effective to perform fine-grained recognition, as can be seen by the good performance on CUB. We owe the success to two aspects. First, the discriminative property of contrastive learning ensures the contrastive network to effectively discriminate the fine-grained classes. Second, the fine-grained images are more effective to transfer the knowledge since the classes are similar, which makes our model more robust to the target classes. A little lower performance is obtained on APY probably due to the weak relations between the source and target classes. APY is a small-scale coarse-grained dataset, where the categories are very different. Therefore, the relations between source and target classes are weak. That's why most approaches could not perform well on this simple dataset. Since we utilize the class similarities to transfer the knowledge, our model may

[1] Source code is available at *http://vipl.ict.ac.cn/resources/codes*.

Method	APY			AWA1			AWA2			CUB			SUN		
	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
DAP [19]	4.8	78.3	9.0	0.0	88.7	0.0	0.0	84.7	0.0	1.7	67.9	3.3	4.2	25.1	7.2
IAP [19]	5.7	65.6	10.4	2.1	78.2	4.1	0.9	87.6	1.8	0.2	72.8	0.4	1.0	37.8	1.8
CONSE [24]	0.0	91.2	0.0	0.4	88.6	0.8	0.5	90.6	1.0	1.6	72.2	3.1	6.8	39.9	11.6
CMT [31]	1.4	85.2	2.8	0.9	87.6	1.8	0.5	90.0	1.0	7.2	49.8	12.6	8.1	21.8	11.8
SSE [41]	0.2	78.9	0.4	7.0	80.5	12.9	8.1	82.5	14.8	8.5	46.9	14.4	2.1	36.4	4.0
LATEM [36]	0.1	73.0	0.2	7.3	71.7	13.3	11.5	77.3	20.0	15.2	57.3	24.0	14.7	28.8	19.5
ALE [1]	4.6	73.7	8.7	16.8	76.1	27.5	14.0	81.8	23.9	23.7	62.8	34.4	21.8	33.1	26.3
DEVISE [10]	4.9	76.9	9.2	13.4	68.7	22.4	17.1	74.7	27.8	23.8	53.0	32.8	16.9	27.4	20.9
SJE [2]	3.7	55.7	6.9	11.3	74.6	19.6	8.0	73.9	14.4	23.5	59.2	33.6	14.1	30.5	19.8
EZSL [27]	2.4	70.1	4.6	6.6	75.6	12.1	5.9	77.8	11.0	12.6	63.8	21.0	11.0	27.9	15.8
SYNC [4]	7.4	66.3	13.3	8.9	87.3	16.2	10.0	90.5	18.0	11.5	70.9	19.8	7.9	43.3	13.4
SAE [17]	0.4	80.9	0.9	1.8	77.1	3.5	1.1	82.2	2.2	7.8	54.0	13.6	8.8	18.0	11.8
CDL [14]	19.8	48.6	28.1	28.1	73.5	40.6	-	-	-	23.5	55.2	32.9	21.5	34.7	26.5
RNet [32]	-	-	-	31.4	91.3	46.7	30.0	93.4	45.3	38.1	61.4	47.0	-	-	-
FGN [38]	-	-	-	57.9	61.4	59.6	-	-	-	43.7	57.7	49.7	42.6	36.6	**39.4**
GAZSL [43]	14.2	78.6	24.0	29.6	84.2	43.8	35.4	86.9	50.3	31.7	61.3	41.8	22.1	39.3	28.3
DCN [20]	14.2	75.0	23.9	25.5	84.2	39.1	-	-	-	28.4	60.7	38.7	25.5	37.0	30.2
TCN (ours)	24.1	64.0	**35.1**	49.4	76.5	**60.0**	61.2	65.8	**63.4**	52.6	52.0	**52.3**	31.2	37.3	34.0

Table 3. GZSL results on APY, AWA1, AWA2, CUB and SUN. ts = Top-1 accuracy of the target classes, tr = Top-1 accuracy of the source classes, H = harmonic mean. We measure average per-class top-1 accuracy in %. '-' represents that the results are not reported.

be influenced by the weak relations.

We argue that traditional approaches usually tend to overfit the source classes since they ignore the target in the model learning process, which will result in the projection domain shift problem. While TCN could alleviate this problem since our model explicitly transfers the knowledge. To demonstrate this viewpoint, we perform GZSL task on these datasets. Table 3 shows the comparison results, where 'ts' is average per-class top-1 accuracy of target classes and 'tr' is the same evaluation results on source classes. 'H' is the harmonic mean that evaluates the total performance. It can be seen that most approaches achieve very high performance on the source classes and extremely low performance on the target classes, which indicates that these approaches learn little knowledge about the target classes. Compared with the results in Table 2, the performance of target classes drops greatly for GZSL because most target-class images are recognized as source classes. This indicates that previous approaches are easy to overfit the source classes. While TCN can effectively alleviate the overfitting problem, as can be seen by the more balanced performance on source and target classes for our approach. We owe the success to the transferable property of the contrastive network, which makes our model more robust to recognize the target images. Although the generative approaches [38, 43] are also very effective in GZSL, they need to learn the complicated generative models. While our approach is very simple to learn. Moreover, our approach is well complementary to the generative approaches since the generated features can also be utilized to learn our model. Some other approaches [14, 20] also adapt the models to the target classes. Compared with them, our approach is more effective.

Dataset	**Method**	**ZSL**	**GZSL**		
			ts	**tr**	**H**
APY	Base	37.52	5.50	77.78	10.28
	TCN	38.93	24.13	64.00	35.05
AWA1	Base	70.15	9.22	64.78	16.14
	TCN	70.34	49.40	76.48	60.03
AWA2	Base	68.48	9.32	54.23	15.91
	TCN	71.18	61.20	65.83	63.43
CUB	Base	56.62	24.70	64.90	37.84
	TCN	59.54	52.58	52.03	52.30
SUN	Base	61.04	21.94	38.64	27.99
	TCN	61.53	31.18	37.29	33.96

Table 4. Comparison with the baseline approach where the knowledge transfer item ($\mathcal{L}_T$) is removed. 'Base' represents the baseline approach. 'TCN' is our approach. 'ZSL' is the accuracy of zero-shot recognition. 'ts', 'tr' and 'H' are the target-class accuracy, source-class accuracy and harmonic mean in GZSL.

We also tried other information fusion approaches and more details are shown in the supplementary materials.

4.3. Importance of Knowledge Transfer

Explicit knowledge transfer is an important part of our framework. It is intuitive that similar objects should play a more important role in transfer learning. Therefore, we use class similarities to explicitly transfer the knowledge from source images to similar target classes. In this way, our model will be more robust to the target classes. More-

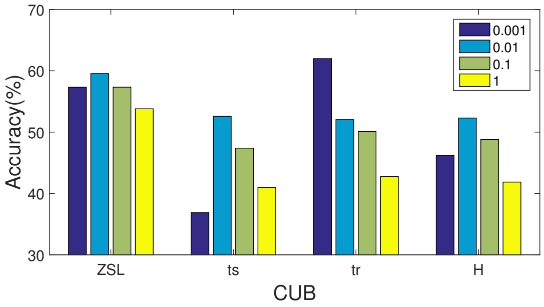

Figure 3. The recognition results on CUB with different value of α. 'ZSL' is the accuracy of zero-shot recognition. 'ts', 'tr' and 'H' are the target-class accuracy, source-class accuracy and harmonic mean in GZSL.

over, it should also have the ability to prevent the model from overfitting the source classes. To demonstrate these assumptions, we compare our approach with the basic model, where the knowledge transfer term ($\mathcal{L}_T$) is removed. Table 4 shows the recognition results. Although only small improvements are achieved for ZSL, the improvements for GZSL are significant. This phenomenon demonstrates that explicit knowledge transfer can effectively tackle the overfitting problem, which enables the model to learn the knowledge about the target classes.

Another factor that deserves to be explored is how important the knowledge transfer is. Therefore, we analyze the influence of parameter α to our model and the recognition results on CUB are shown in Figure 3. It can be seen that TCN achieves its best performance when α equals to 0.01. We can infer that α should be small in order to get good performance. This may be caused by two reasons. First, the class similarities are fuzzy measures and there is no accurate definitions. Second, the source images do not absolutely match with the target classes. When α increases, the performance of source classes drops, as can be seen by the results of 'tr', because the model pays more attention to the target classes and neglects the accurate source classes. Since the loss on the source classes ensures the discriminative property of contrastive learning and the loss on the target classes guarantees the transferable property, we must balance these terms to obtain a robust recognition model.

4.4. Visualization of Class Similarities

The transferable property of our approach is accomplished by leveraging the class similarities to make knowledge transfer in the model learning process. To see what knowledge has been transferred, we show the class similarities of AWA1 in Figure 4. Because of space constraints, we select 15 source classes and visualize their similarities to the target classes. It can be figured out that *leopard* is similar to *bobcat* so the training samples of *leopard* can also be utilized to learn the target class *bobcat* in the training phase, thus to enable knowledge transfer. It effectively tackles the problem that no training images are available for the target classes. Through such explicit knowledge transfer, our model would be more robust to the target. Other class similarities, *i.e. killer+whale* is similar to *blue+whale*, *seal*, *walrus* and *dolphin*, are also useful knowledge to transfer in the contrastive learning process.

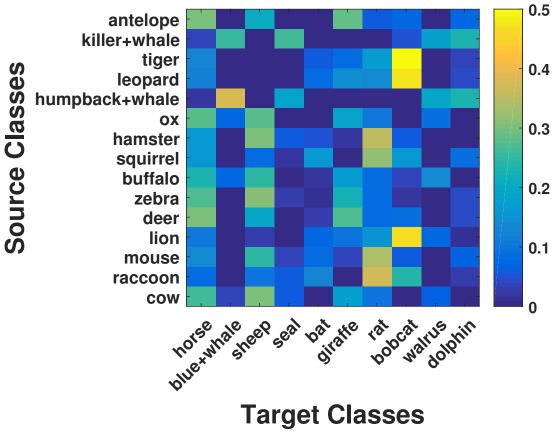

Figure 4. The class similarities in AWA1, where 15 source classes are selected. Each row represents the similarities of one source class to the target classes.

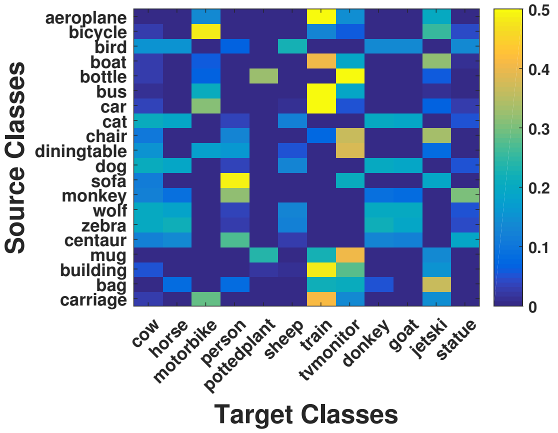

Figure 5. The class similarities in APY, where each row shows the similarities of one source class to the target classes.

The foundation on which our approach works well is that reasonable class similarities are obtained for knowledge transfer. However, the class similarities may be very rough for some coarse-grained dataset, such as APY, so it becomes difficult to transfer knowledge from source classes to the target classes. That is why low zero-shot recognition accuracy is obtained on APY for all approaches, as can be seen from Table 2. To make it intuitive, we show the class similarities for APY in Figure 5. It can be figured out that the relations between source and target classes are less reli-

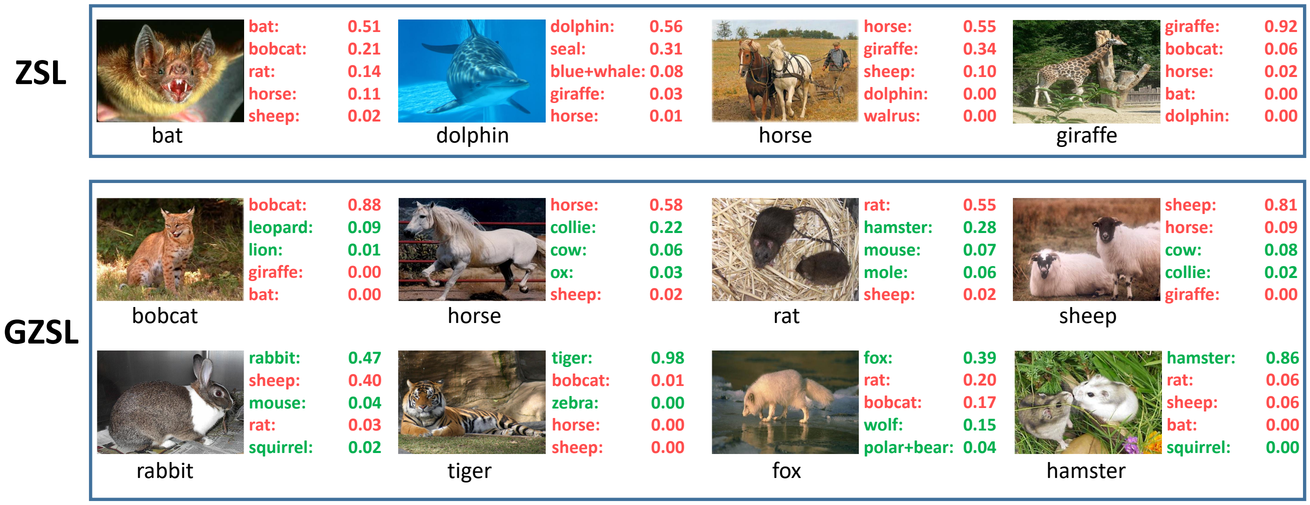

Figure 6. The normalized contrastive values of some test samples obtained on AWA1, where five most similar classes are shown. The source classes are marked with green and the target classes are marked with red. The first row in GZSL shows the target-class samples and the second shows the source-class samples.

able. For example, among the target classes, the most similar one to the source class *building* is the *train*. However, buildings and trains are very different in reality. Therefore, using the training images of *building* to learn the target *train* would degrade our model. This may be the reason why TCN achieves lower performance than the state-of-the-art approach on APY. Although some incomprehensible similarities exist, there are also some useful relations, *i.e. bicycle* is similar to *motorbike* and *bus* is similar to *train*, which ensures the relative good performance of our approach.

4.5. Visualization of Contrastive Values

Different from the visual-semantic embedding approaches that use fixed distance to conduct zero-shot recognition, our transferable contrastive network automatically contrasts one image with every class and outputs the contrastive values for image recognition. Figure 6 shows the contrastive values of some test samples obtained on AWA1. In order to make it intuitive, we normalize the contrastive values and show five most similar classes, where the target classes are marked with red and the source classes are marked with green. We can figure out that most images are consistent with their corresponding classes and dissimilar with other classes. For ZSL, we recognize the test samples among the target classes. As can be seen, the image 'giraffe' has high contrastive value with its class and has low contrastive values with other ones. For GZSL, we recognize the test samples among all classes. Although we only have source-class images for training, our model can effectively recognize the target-class samples in the test procedure. For example, 'bobcat' is effectively discriminated with source-class *leopard* in GZSL task though these two classes are very similar. We owe this success to the explicit knowledge transfer by the class similarities. It prevents our model from overfitting the source classes and ensures the transferable ability to target classes, thus the target-class images would be effectively recognized when they are encountered. Moreover, one image may also have relatively high contrastive values with similar classes. For example, 'rat' has relative strong activations on *hamster*. This shows that TCN is not only discriminative enough to classify different classes but also transferable to novel classes.

5. Conclusion

In this paper, we propose a novel transferable contrastive network for generalized zero-shot learning. It automatically contrasts the images with the class semantics to judge how consistent they are. We consider two key properties in contrastive learning, where the discriminative property ensures the contrastive network to effectively classify different classes and the transferable property makes the contrastive network more robust to the target classes. By explicitly transferring knowledge from source images to similar target classes, our approach can effectively tackle the problem of overfitting the source classes in GZSL task. Extensive experiments on five benchmark datasets show the superiority of the proposed approach.

Acknowledgements. This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos.61390511, 61772500, CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009, and Youth Innovation Promotion Association No. 2015085.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 819–826, 2013.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[3] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. In *Proc. of European Conference on Computer Vision*, pages 730–746, 2016.

[4] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[5] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *Proc. of International Conference on Computer Vision*, pages 3496–3505, 2017.

[6] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proc. of European Conference on Computer Vision*, pages 52–68, 2016.

[7] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding network. In *Proc. of Computer Vision and Pattern Recognition*, pages 1043–1052, 2018.

[8] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *Proc. of International Conference on Learning Representations workshops*, 2015.

[9] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proc. of Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, MarcAurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proc. of Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[11] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2332–2345, 2015.

[12] Zhen-Yong Fu, Tao A. Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proc. of Computer Vision and Pattern Recognition*, pages 2635–2644, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *Proc. of European Conference on Computer Vision*, pages 118–134, 2018.

[15] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Adaptive metric learning for zero-shot recognition. *Signal Processing Letters*, 26(9):1270–1274, 2019.

[16] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *Proc. of International Conference on Computer Vision*, pages 4233–4242, 2017.

[17] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 4447–4456, 2017.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[19] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. of Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[20] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Generalized zero-shot learning with deep calibration network. In *Proc. of Advances in Neural Information Processing Systems*, pages 2005–2015, 2018.

[21] Yang Long, Li Liu, Fumin Shen, Ling Shao, and Xuelong Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2498–2512, 2018.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[23] Pedro Morgado and Nuno Vasconcelos. Semantically consistent regularization for zero-shot recognition. In *Proc. of Computer Vision and Pattern Recognition*, pages 2037–2046, 2017.

[24] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Gregory S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proc. of International Conference on Learning Representations*, 2014.

[25] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The SUN attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[26] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proc. of Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[27] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *Proc. of International Conference on Machine Learning*, pages 2152–2161, 2015.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151, 2015.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[31] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Proc. of Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[32] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. of Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[35] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proc. of Computer Vision and Pattern Recognition*, pages 6857–6866, 2018.

[36] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[37] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019.

[38] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 5542–5551, 2018.

[39] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *Proc. of Computer Vision and Pattern Recognition*, pages 3077–3086, 2017.

[40] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 3010–3019, 2017.

[41] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proc. of International Conference on Computer Vision*, pages 4166–4174, 2015.

[42] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proc. of Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.

[43] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed M. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proc. of Computer Vision and Pattern Recognition*, pages 1004–1013, 2018.