DEEP META-RELATION NETWORK FOR VISUAL FEW-SHOT LEARNING

Fahong Zhang, Qi Wang, Xuelong Li*

School of Computer Science and Center for OPTical IMagery Analysis and Learning,
Northwestern Polytechnical University, Xi'an, Shaanxi, China

ABSTRACT

This paper proposes a novel metric-based deep learning method to solve the few-shot learning problem. It models the relation between images as high dimensional vector, and trains a network module to judge, when given two relational features, which one indicates a stronger connection between the image objects. By training such a network module, we introduce a comparative mechanism into the metric space, i.e., the similarity score of any two images is computed after seeing other images in the same task. Further more, we propose to incorporate a batch classification loss into episodic training to mitigate the hard training problem that occurs when embedding network is going deeper. Experiments demonstrate that the proposed network can achieve promising performance.

***Index Terms*—** Few-shot learning, deep learning, metric learning

1. INTRODUCTION

It is always a meaningful and interesting problem that how can machine learning systems obtain human's remarkable ability of learning novel visual concepts with only few examples seen. Generally, the above problem is modeled as few-shot learning in computer vision, in which a learning system is asked to perform N-way classification over query images with K (K is usually less than 10) support images seen in each category. One philosophy to tackle few-shot learning problem is metric-learning [1, 2, 3, 4, 5]. It aims to find a metric space in which instances' projection have large inter- and small intra-class margin. Prior works in this type usually use convolutional neural network (CNN) [6, 7] to extract visional feature, and design different distant metric to guide the network in searching for the perfect metric space.

However, a perfect metric space is always hard to be found, because relation between images is multi-dimensional and hard to be quantified. For example, in Fig. 1, one can observe that the query image is more similar to image 1 in color and orientation, while to image 2 in shape and texture. That indicates images belong to different categories may also share strong similarity. So why can human make correct prediction towards this task so easily? Our assumption is that human are not only expert in capturing and organizing the visional characteristics of the images, but also good at modeling their relation and *distinguishing which type of relation is more significant*. In this example, images' similarity in shape and texture is considered to be more important according to most of our knowledge, and this leads to the conclusion that the query instance is more likely to be a ginkgo leaf rather than a maple leaf.

Fig. 1. An example of one-shot learning problem.

Motivated by the above observation, this paper proposes a Meta-Relation Network (MRN) to solve the few-shot learning problem. The contributions are listed as follows.

- We model the relation between images as high dimensional feature (named as relational feature), and design a network architecture to evaluate the meta-relation (or relation of relational feature) among images, i.e., to decide which one indicates a stronger similarity when two relational features are given. In this way, we avoid the difficulty in quantifying the similarity between images as mentioned above.
- We propose to incorporate batch classification loss into episodic training [8] to mitigate the hard training problem [9, 10] which usually occurs when feature extraction network is going deeper.

2. RELATED WORK

The existing methods in the literature of few-shot learning can be roughly categorized into optimization-based, memory-

*Qi Wang is the corresponding author. This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470, and 61761130079.

based and metric learning-based approaches.

Optimization-based approaches [11, 12] perform parameter adaption for a meta-learned network model towards the novel task. The main problems they are facing are, how to ensure the adaption is fast enough, and will not lead to overfitting in novel task or forgetting of previous knowledge. MAML [13] is proposed with the core idea to accelerate the adaption of parameters. It trains the network in a way that the network can quickly make adaption to the simulated task sampled from training set, so that the trained network can potentially make faster adaption on novel task.

Memory-based methods [14] stores the past knowledge in specific form, e.g., cell state of RNN-based architecture or images' representation in external memory. They try to build the connection between novel task images and these historic information so as to guide the network to make better decision. For example, SNAIL [15] process the task images in a sequential manner. It accesses the past knowledge by performing temporal convolution over a pre-defined length of previous images' representation.

Metric learning-based approaches [16] accomplish few-shot learning by measure the similarity between task images. Specifically, they aim to find the most similar images to the query image in support set by learning a metric space where images belong to the same class are close to each other. Matching Network [8] firstly applies the cosine distance to measure the similarity between two images in feature space. Relation Network [17] proposes to use a stack of trainable fully connected layers to serve as the distance metric.

3. METHODOLOGY

3.1. Problem Formulation

In few-shot classification, we are first given two dataset $\mathcal{D}_{train} = \{(\mathbf{x}_i, y_i) \mid y_i \in \mathcal{Y}_{train}\}_{i=1}^{M_{train}}$ and $\mathcal{D}_{test} = \{(\mathbf{x}_i, y_i) \mid y_i \in \mathcal{Y}_{test}\}_{i=1}^{M_{test}}$, where $\mathcal{Y}_{train}$ and $\mathcal{Y}_{test}$ are training and testing label sets disjoint with each other. M_{train} and M_{test} are dataset sizes. $\mathbf{x}_i$ denotes an image in dataset and y_i is its label. We are then supposed to train a learning system using $\mathcal{D}_{train}$ and optimize its performance on task set $\mathcal{T} = \{\tau_i\}$ generated from $\mathcal{D}_{test}$. Here a single task $\tau \in \mathcal{T}$ consists of a *support set* $\mathcal{S}$ and a *query set* $\mathcal{Q}$: $\tau = (\mathcal{S}, \mathcal{Q})$. For N-way K-shot learning, $\mathcal{S} = \{\{\overline{\mathbf{x}}_{i,j}\}_{j=1}^{K}\}_{i=1}^{N}$ consists of N categories of images sampled from $\mathcal{D}_{test}$. Each category contains K samples. $\mathcal{Q} = \{\tilde{\mathbf{x}}_i\}_{i=1}^{q}$ has q images sampled from $\mathcal{D}_{test}$, and they share the same label space with $\mathcal{S}$. In convenience, we only consider the case that $q = 1$ (so we can refer $\tilde{\mathbf{x}}$ to the query image in the following sections). In this case, there will be K images in $\mathcal{S}$ that have the same label with $\tilde{\mathbf{x}}$. They consist a *interest set* $\mathcal{I} = \{\overline{\mathbf{x}}_{\tilde{i},j}\}_{j=1}^{K} \in \mathcal{S}$, where $\tilde{i}$ indexes the label of $\tilde{\mathbf{x}}$ out of N different categories.

When referring to how to train a learning system using $\mathcal{D}_{train}$, [8] proposes a episodic training strategy and argues that training procedure should match the inference at test time. We follow this strategy and hence the data input to the network will always be in task form, i.e., containing a support set $\mathcal{S}$ and a query set $\mathcal{Q}$.

3.2. Network Workflow

The overall architecture of the proposed MRN is shown in Fig. 2. MRN is composed of three modules named encoding, relation and meta-relation module.

Encoding module f_{en} is a general CNN-based feature encoder. When receiving an input task τ, it maps the images in $\mathcal{S}$ and $\mathcal{Q}$ into feature embeddings.

The design of relation module is inspired by [17]. It first concatenates each of the NK support image embeddings with the query image embedding individually, and then maps the concatenated features into lower dimensional vectors through a stack of fully connected (FC) layers f_{re}. These vectors, each corresponds to a support image, are named as relational features as they represent the relation between each of the support images and the query image. We denote them as $\mathcal{R} = \{\{\mathbf{r}_{i,j}\}_{j=1}^{K}\}_{i=1}^{N}$. The only difference between our relation module and the structure in [17] is that, we calculate a relational feature for each support image, but [17] first averages support image embeddings that belong to the same category and hence produce only N relational features eventually.

Meta-relation module is built to measure the relation between relational features, i.e., to judge which one is more decisive w.r.t. the classification task when given two relational features. Specifically, relational features for the same class are first averaged to reduce the data size. As a result, we obtain N averaged relational features $\{\overline{\mathbf{r}}_i\}_{i=1}^{N}$. Then, elements in $\{\overline{\mathbf{r}}_i\}_{i=1}^{N}$ are concatenated pair-wisely to generate N^2 pairs of combined features. Each of them is feeded to another stack of FC layers (denoted by f_{me}) to produce a meta-relation score. After this operation we can get a probabilistic matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ with N^2 scores. Note that each element $m_{i,j}$ in $\mathbf{M}$ represents the relation between $\overline{\mathbf{r}}_i$ and $\overline{\mathbf{r}}_j$, i.e., the probability that $\overline{\mathbf{r}}_i$ reflects a stronger connection than $\overline{\mathbf{r}}_j$. Finally, by calculating the sum of each row of $\mathbf{M}$, we can achieve a classification score $\mathbf{c} \in \mathbb{R}^N$ to predict the category of the query image.

3.3. Loss Definition

This subsection will describe the loss functions used to train MRN. There are four losses in total.

Batch classification loss L_{ba}. The purpose to apply this loss is to mitigate the hard training problem mentioned in section 1. For each task $\tau = (\mathcal{S}, \mathcal{Q}) \in \mathcal{T}$ where $\tilde{\mathbf{x}} \in \mathcal{Q}$ is the query image, the batch classification loss is calculated by:

$$L_{ba} = \sum_{\tau \in \mathcal{T}} \mathcal{L}_{ce}(f_l^{(1)}(f_{en}(\tilde{\mathbf{x}})), \tilde{y}). \tag{1}$$

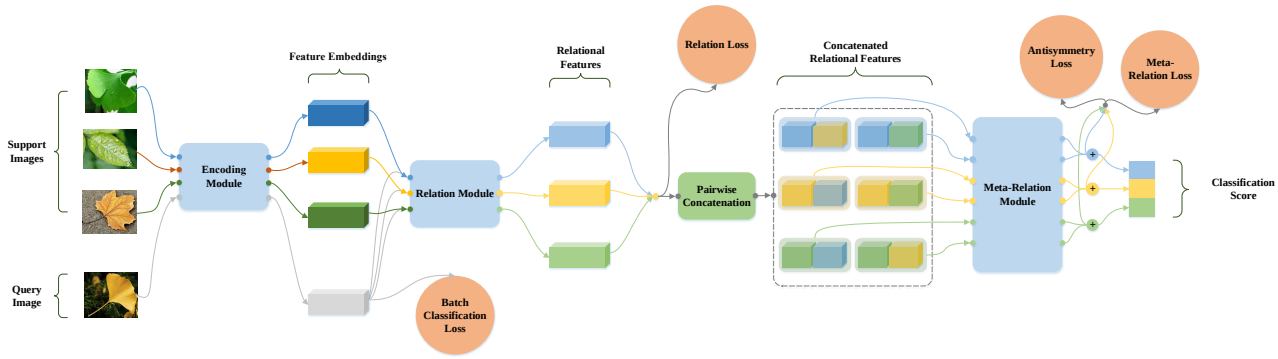

Fig. 2. The overall architecture of MRN.

Here $\mathcal{L}_{ce}(\cdot)$ is the commonly used cross-entropy loss function, $\tilde{y}$ is the label of $\tilde{\mathbf{x}}$, and $f_l^{(1)}$ is a fully connected layer which maps the feature embeddings extracted by f_{en} to a $|\mathcal{Y}_{train}|$-dimensional classification score ($|\cdot|$ is the cardinality of a set).

Relation loss L_{re}. L_{re} is applied to guide the relation module to better metric the relation between visional features. When averaged relational features $\{\bar{\mathbf{r}}_i\}_{i=1}^N$ are received from relation module, they are mapped into a classification scores $\mathbf{z} \in \mathbb{R}^N$ through a stack of FC layer $f_l^{(2)}$: $\mathbf{z} = [f_l^{(2)}(\bar{\mathbf{r}}_1), f_l^{(2)}(\bar{\mathbf{r}}_2), ..., f_l^{(2)}(\bar{\mathbf{r}}_N)]$. Then L_{re} can be calculated as:

$$L_{re} = \sum_{\tau \in \mathcal{T}} \mathcal{L}_{ce}(\mathbf{z}, \tilde{i}), \tag{2}$$

Meta-relation loss L_{me}. L_{me} supervises the meta-relation module in judging the relationship of two relational features. The core idea is that for images in interest set $\mathcal{I}$, their corresponding relational features should dominate the others. More formally, elements in row $\tilde{i}$ of $\mathbf{M}$ should have higher values while elements in col $\tilde{i}$ should have lower values. In consideration of this, L_{me} is defined as:

$$L_{me} = \sum_{\tau \in \mathcal{T}} \sum_{j=1}^{n} \|m_{\tilde{i},j} - \epsilon_{high}\|_2 + \sum_{j=1}^{n} \|m_{j,\tilde{i}} - \epsilon_{low}\|_2. \tag{3}$$

Here ϵ_{high} and ϵ_{low} are a pair of relatively "high" and "low" values. They are set to 0.8 and 0.2 empirically.

Antisymmetry Loss L_{an}. Another consideration is that $\mathbf{M}$ should be antisymmetric in probability sense. When given two relational features $\bar{\mathbf{r}}_i$ and $\bar{\mathbf{r}}_j$, if $\bar{\mathbf{r}}_i$ indicates a stronger relation than $\bar{\mathbf{r}}_j$ with possibility $m_{i,j}$, $\bar{\mathbf{r}}_j$ should be stronger than $\bar{\mathbf{r}}_i$ with possibility $1 - m_{i,j}$. That means the sum of $m_{i,j}$ and $m_{j,i}$ should be close to 1. In consideration of this, $\mathcal{L}_{an}$ is defined as:

$$L_{an} = \sum_{\tau \in \mathcal{T}} \|\mathbf{M} + \mathbf{M}^T - \mathbf{1}_{n \times n}\|_F, \tag{4}$$

where $\mathbf{1}_{n \times n}$ is a $n \times n$ matrix filled with 1 and $\|\cdot\|_F$ is Frobenius norm.

Finally, our goal is to minimize a linear combination of the above four loss functions.

$$\min_{f_{en}, f_{re}, f_l^{(1)}, f_{me}, f_l^{(2)}} L_{ba} + \alpha L_{re} + \beta L_{me} + \gamma L_{an}. \tag{5}$$

4. EXPERIMENTS

4.1. Dataset

The evaluation dataset used in the experiments is *mini*ImageNet, a subset of ILSVRC-12 dataset. It consists of $60,000$ images sampled from 100 categories in ImageNet, and each category has 600 images. In this paper we adopt the split provided by [18]. It contains 64 categories for training, 16 for validation and 20 for testing. All the images are resized to 84×84 and augmented by random reflecting, random cropping and color jittering.

4.2. Network architecture

The whole MRN consists of three modules: encoding module, relation module and meta-relation module. For encoding module, we adopt a 40-layer Wide Residual Network (WRN) [19] with widen factor 4. Considering that feature from the laster layer may be too class-specific and lack generalization ability, the feature output by the second block is chosen to be the feature vector. As introduced in Section 3.3, the output of the third block is used to calculate the batch classification loss. The architecture of relation module and meta-relation module are shown in Fig. 3. They consist of stacked fully connected layers and ReLU or sigmoid unit.

4.3. Training details

We adopt the episodic training strategy proposed in [8], where training data has exactly the same form as testing data. In our

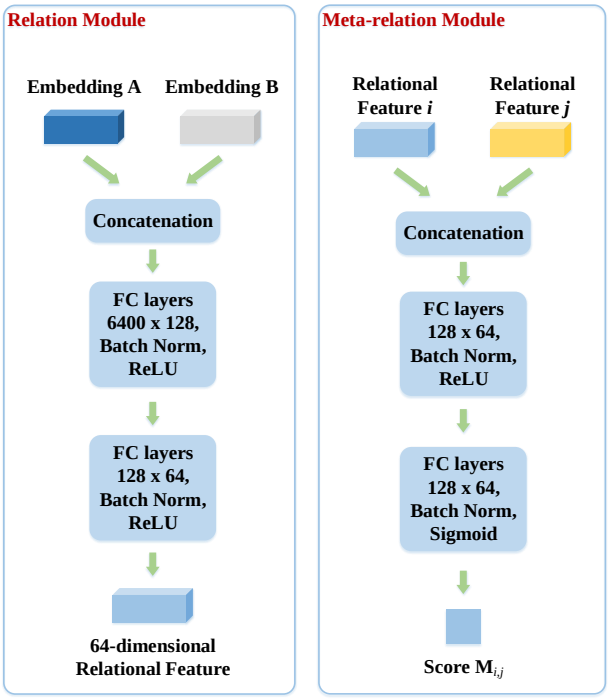

Fig. 3. Architecture of relation and meta-relation module.

setting, each input batch contains 2 tasks. For N-way K-shot learning, the support set in each task is sampled from N different classes with K images in each class. For query set, the number of query images for each class is set to 5. SGD with nesterov [20] is selected as the optimizer. The learning rate of it is set to 0.1 initially and multiplied by 0.8 after every 8000 batches.

Our proposed MRN involves 3 hyper-parameters α, β and γ in Eq. (5). They are set to $\alpha = \beta = \gamma = 0.5$ by tuning on validation set. We found that network's performance is not very sensitive to them.

4.4. Experimental Results

Table 1 shows the classification accuracy of MRN and some comparative methods on *mini*ImageNet dataset. According to the result, the proposed MRN achieves the best performance on 5-way 1-shot task. While on 5-way 5-shot task, MRN can also achieve a satisfactory performance, with only 0.3% loss comparing to the best method.

4.5. Ablation Study

The major novelty of this paper falls on the incorporation of meta-relation and batch classification loss. So in this section, some ablation studies are conducted to verify their effectiveness.

Method	5-way Acc.	
	1-shot	5-shot
Matching Net [1]	$43.56 \pm 0.84\%$	$55.31 \pm 0.73\%$
Proto Net [1]	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$
Relation Net [17]	$50.44 \pm 0.82\%$	$65.32 \pm 0.70\%$
k-shot [9]	$56.3 \pm 0.4\%$	$\mathbf{73.9 \pm 0.3\%}$
adaResNet [9]	$56.88 \pm 0.62\%$	$71.94 \pm 0.57\%$
SNAIL [10]	$55.71 \pm 0.99\%$	$68.88 \pm 0.92\%$
PFA [21]	$59.60 \pm 0.41\%$	$73.74 \pm 0.19\%$
MRN (ours)	$\mathbf{61.14 \pm 0.57\%}$	$73.68 \pm 0.45\%$

Table 1. Few-shot classification results on *mini*ImageNet dataset. Each accuracy result is reported with 95% confidence interval. Best result in each task is highlighted. The first three methods use shallower network as feature encoder, while the others adopt deeper architecture like ResNet and WRN.

L_{me}	L_{ba}	5-way Acc.	
		1-shot	5-shot
Y	Y	$\mathbf{61.14 \pm 0.57\%}$	$\mathbf{73.68 \pm 0.45\%}$
Y	N	$55.64 \pm 0.60\%$	$70.96 \pm 0.47\%$
N	Y	$59.56 \pm 0.48\%$	$72.45 \pm 0.44\%$
N	N	$53.77 \pm 0.59\%$	$70.24 \pm 0.49\%$

Table 2. Classification accuracy comparison on *mini*ImageNet dataset under whether meta-relation loss L_{me} and batch classification loss L_{ba} are used or not. Accuracy is reported with 95% confidence interval.

Based on whether these two losses are used to train the network, four comparative results are reported in Table 2. As shown in the table, both meta-relation loss and batch classification loss can improve the network's performance. Especially for batch classification loss, it helps to improve the 5-way 1-shot accuracy by more than 5%.

5. CONCLUSION

In this paper, a novel meta-relation network is proposed to solve the few-shot classification problem. Based on the assumption that the relation between two concerned images is complex and multi-dimensional, a deep neural network is learned to judge which one indicates a stronger connection between images when two relational features are given. Moreover, in order to mitigate the problem that deeper feature extraction network is hard to be trained, batch classification loss L_{ba} is proposed to accelerate the training procedure. Experiments show that it not only makes the network end-to-end trainable, but also largely improves its performance.

6. REFERENCES

[1] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[2] Szu-Yu Chou, Kai-Hsiang Cheng, Jyh-Shing Roger Jang, and Yi-Hsuan Yang, "Learning to match transient sound events using attentional similarity for few-shot sound recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 26–30.

[3] Jixuan Wang, Kuan Chieh Wang, Marc Law, Frank Rudzicz, and Michael Brudno, "Centroid-based deep metric learning for speaker recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3652–3656.

[4] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 46–58, 2020.

[5] Qi Wang, Junyu Gao, and Xuelong Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4376–4386, 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[9] Matthias Bauer, Mateo Rojas-Carulla, Jakub Bartlomiej Swiatkowski, Bernhard Schölkopf, and Richard E. Turner, "Discriminative k-shot learning using probabilistic models," *arXiv preprint arXiv:1706.00326*, 2017.

[10] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Advances in Neural Information Processing Systems*, 2018, pp. 719–729.

[11] Yoonho Lee and Seungjin Choi, "Gradient-based meta-learning with learned layerwise metric and subspace," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2927–2936.

[12] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas, "Learning to learn by gradient descent by gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 3981–3989.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1126–1135.

[14] Tsendsuren Munkhdalai and Hong Yu, "Meta networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 2554–2563.

[15] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel, "A simple neural attentive meta-learner," in *International Conference on Learning Representations*, 2018.

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015, vol. 2.

[17] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018*, 2018, pp. 1199–1208.

[18] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings*, 2017.

[19] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *CoRR*, vol. abs/1605.07146, 2016.

[20] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning, ICML*, 2013, pp. 1139–1147.

[21] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.