SGAP-Net: Semantic-Guided Attentive Prototypes Network for Few-Shot Human-Object Interaction Recognition

Zhong Ji,[1] **Xiyao Liu,**[1] **Yanwei Pang,**[1*] **Xuelong Li**[2]
[1]Tianjin University, [2]Northwestern Polytechnical University
{jizhong, xiyaoliu, pyw}@tju.edu.cn, xuelong_li@ieee.org

Abstract

Extreme instance imbalance among categories and combinatorial explosion make the recognition of Human-Object Interaction (HOI) a challenging task. Few studies have addressed both challenges directly. Motivated by the success of few-shot learning that learns a robust model from a few instances, we formulate HOI as a few-shot task in a meta-learning framework to alleviate the above challenges. Due to the fact that the intrinsic characteristic of HOI is diverse and interactive, we propose a Semantic-Guided Attentive Prototypes Network (SGAP-Net) to learn a semantic-guided metric space where HOI recognition can be performed by computing distances to attentive prototypes of each class. Specifically, the model generates attentive prototypes guided by the category names of actions and objects, which highlight the commonalities of images from the same class in HOI. In addition, we design a novel decision method to alleviate the biases produced by different patterns of the same action in HOI. Finally, in order to realize the task of few-shot HOI, we reorganize two HOI benchmark datasets, i.e., HICO-FS and TUHOI-FS, to realize the task of few-shot HOI. Extensive experimental results on both datasets have demonstrated the effectiveness of our proposed SGAP-Net approach.

Introduction

Human-Object Interaction (HOI) is to recognize the relationship between human actions and surrounding environment, which is a challenging computer vision task. Unlike object recognition, the key to understanding visual scene is not only to recognize the existence of instances, but also to understand the visual relationship among instances. Instead of "Where is what" (i.e., object detection and recognition in images), the goal of HOI recognition is to answer "What does the person do with something". That is to say, given an input image, this task aims at recognizing the triple $< human, verb, object >$.

Despite significant advances in the field of computer vision, such as image classification (He et al. 2016; Wang et al. 2017), object detection (Ren et al. 2017; He et al. 2017) and action recognition (Gkioxari, Girshick, and Malik 2015; Girdhar and Ramanan 2017), the study of HOI is still in its infancy. This is due to: (1) The instance imbalance among categories increases the difficulty of training and is prone to cause over-fitting (Chao et al. 2015). Some common interactions have abundant samples, while some unusual HOI combinations have quite few samples. For example, samples of "Hold-Horse" are much more than those of "Feed-Horse", as shown in Fig. 1. (2) Multiple labels in HOI make the number of classes increase exponentially, which results in the various information being entangled with each other, or aggregating the similarity of all possible pairs of actions and objects exhaustively, and bringing about huge burden for computing resources. Current approaches apply data augmentation with weakly labeled data (Zhuang et al. 2018) or valid geometric configuration regularization (Lin et al. 2016) to improve performance for HOI recognition. However, they neglect the above challenges in HOI.

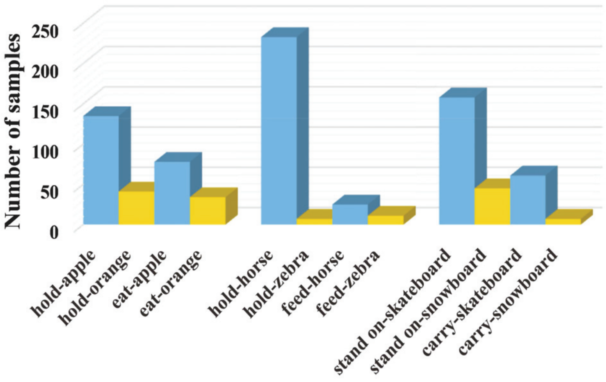


Figure 1: Extreme imbalance and ubiquitous similarities among categories on one HOI dataset: HICO (Chao et al. 2015). Despite of the same object category, the number of samples produced by different actions varies greatly. On the other side, objects with similar attributes (e.g., horse and zebra) are likely to interact with the same action.

Recently, the task of few-shot learning has been proved to alleviate the instance imbalance problem effectively. It learns a network that maps an unlabeled example to its label from a small labeled support set (Vinyals et al. 2016),

[*]Corresponding author.

which alleviates over-fitting caused by the instance imbalance problem. Moreover, the idea of meta-learning provides a new solution to combinatorial explosion in HOI. When collecting more new combinations of actions and objects, it can borrow some knowledge from existing HOI models instead of training a new model from scratch due to the ubiquitous similarities among actions. For example, the actions with "Horse" are helpful for recognizing the interactions between humans and "Zebra" (as shown in Fig. 1). To this end, we explore formulating HOI as a few-shot HOI task and applying the idea of meta-learning to address it.

To overcome the diversity and interactivity characteristics in HOI, we propose a semantic-guided attentive prototypes network for few-shot HOI. It includes a SemantIc-Guided Multiple Attention Module (SIGMA-Module) and a Prototypes Shift Module (PS-Module), respectively. Due to the characteristics of HOI scene, the SIGMA-Module is designed to generate class prototypes with multiple attention corresponding actions and objects. In the PS-Module, we design different decision methods for training and testing. Particularly, we employ a super-parameter regulating shift distance in training to alleviate the offset of different patterns produced by the same action.

It is worthwhile to highlight several aspects of the proposed approach here:

- We formulate HOI recognition as a few-shot learning task in a meta-learning framework to alleviate the challenges of instances imbalance and combinatorial explosion. To the best of our knowledge, this is the first work to model HOI recognition into a few-shot learning task.
- We design SGAP-Net to learn a good metric space for HOI, where HOI recognition can be performed by computing distances to attentive class prototypes. Specifically, the SIGMA-Module makes class prototypes focus on discriminative regions of HOI. And the PS-Module is further developed to apply a novel decision method on training to alleviate the offsets of different patterns produced by the same action.
- For evaluating our proposed SGAP-Net in the few-shot HOI recognition, we reorganize two benchmark datasets, HICO-FS and TUHOI-FS. Experiments show that the performance of the proposed SGAP-Net outperforms the state-of-the-art few-shot learning methods in a large margin.

Related work

Human-Object Interaction. Modeling human action and activity has a rich history in computer vision. Recently, the idea of recognizing the interaction between humans and objects has attracted extensive attention (Gupta and Davis 2007). Some studies focused on HOI detection task (Gkioxari et al. 2018; Li et al. 2019b), which applied popular detectors jointly to detect people and objects, and inferred interaction triplets by fusing these predictions. Additionally, compositional learning (Kato, Li, and Gupta 2018) addressed the zero-shot learning problem in HOI. They employed graph convolutional networks to recognize unseen interactions with external knowledge. Moreover, some work focused on creating large scale image datasets for HOI (Le, Uijlings, and Bernardi 2014; Chao et al. 2015). However, current studies neglect the instance imbalance and combinatorial explosion challenges in HOI recognition. Our work takes a step forward by formulating HOI as few-shot HOI task and utilizes meta-learning based few-shot learning for recognizing human-object interactions to alleviate both challenges.

Meta-Learning Based Few-Shot Learning. It is designed to train a meta classifier with limited labeled examples, which transfers the knowledge from seen tasks to unseen tasks. Several recent meta-learning methods (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018) build metric-based networks to realize few-shot learning. For example, Prototypical Networks (Snell, Swersky, and Zemel 2017) learn prototypes for classes and classify the query image into the nearest class prototypes. Other approaches (Finn, Abbeel, and Levine 2017; Nichol, Achiam, and Schulman 2018) propose to learn a good initialization that can be optimized in a few gradient steps and effectively obtain optimal model parameters for new tasks. Different from the above studies, we apply few-shot learning to a novel visual scene, i.e., human-object interaction, which is a challenge for current frameworks. The task of few-shot HOI focuses on learning a robust model with a few samples to recognize visual relationships between humans and novel objects.

Attention Mechanism. The attention mechanism has been widely applied in computer vision and machine learning domain owing to its effectiveness, e.g., image classification (Wang et al. 2017), semantic segmentation (Fu et al. 2018; Wei et al. 2017) and video summarization (Ji et al. 2019). Moreover, many studies (Annadani and Biswas 2018; Yu et al. 2018; Xing et al. 2019) demonstrate that some available prior knowledge can improve performance in zero/few-shot learning as a type of attention information. For example, Yu $et\,al.$ (Yu et al. 2018) proposed a novel stacked semantic-guided attention model to generate attention to the importance of different local regions for fine-grained zero-shot learning. Accordingly, in this work, we design SGAP-Net to generate attentive prototypes relative to the semantic of HOI in the framework of few-shot learning.

Table 1: The notations used in SGAP-Net.

Notations	Description
N	Number of examples in the training set
K	Number of classes in the training set
N_C	Number of classes per episode
N_S	Number of support examples per class
N_Q	Number of query examples per class
e	The e-th episode
c	Class c
S_e^c	Support examples of class c in episode e
Q_e^c	Query examples of class c in episode e
s_i	The i-th support example in S_e^c
q	The query example in Q_e^c

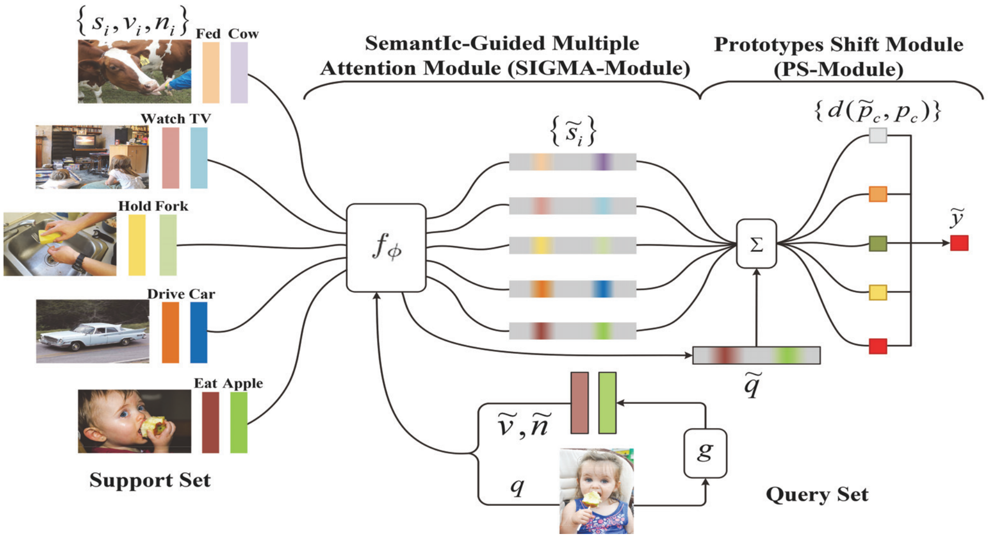


Figure 2: The proposed SGAP-Net architecture for a 5-way 1-shot task. Semantic embeddings of nouns and verbs are given in support set, while the semantic features of query images are generated by the generation network g. f_ϕ is the semantic-guided part of SIGMA-Module, and $\sum$ represents the prototypes calculation part to obtain $\widetilde{p_c}$ and p_c in the PS-Module.

SGAP-Net for Few-Shot HOI Recognition

Problem Definition

We formulate HOI recognition as a few-shot task in a meta-learning framework to alleviate the challenges of instance imbalance and combinatorial explosion in HOI. We follow a C-way K-shot problem using the episodic formulation from Matching Networks (Vinyals et al. 2016). According to the datasets of HOI, the label for an image combines an action and an object. Due to ubiquitous similarities in the same action interacting with different objects and the number of objects are far more than that of actions, we take objects as different tasks in our work. Specifically, we divide all noun labels into the meta-train set, the meta-validation set and the meta-test set that are disjoint in nouns.

In few-shot classification, there are N labeled examples $\{(x_1, y_1), ..., (x_N, y_N)\}$, where each $x_i \in \mathbb{R}^D$ is a D-dimensional visual feature vector of the i-th image, $y_i \in \{1, ..., K\}$ is the corresponding label. When training in the support set, the semantic vectors of labels are given as $\{(n_1, v_1), ..., (n_N, v_N)\}$, where $n_i \in \mathbb{R}^V$ is the V-dimensional text semantic embedding of the noun label, $v_i \in \mathbb{R}^V$ is the V-dimensional text semantic embedding of the verb label. Other notations used in SGAP-Net are listed in Table 1.

The architecture of SGAP-Net consists of a SemantIc-Guided Multiple Attention Module (SIGMA-Module), and a Prototypes Shift Module (PS-Module), as shown in Fig. 2. In the SIGMA-Module, the visual feature representation part is a pretrained ResNet-18 (He et al. 2016). Then we apply the semantics of corresponding verbs and nouns to guide class prototypes with multiple attention, which provides discriminative knowledge of different classes. Finally, the PS-Module helps to learn a good metric space for HOI by a strict decision method.

Semantic-Guided Multiple Attention Module

To obtain an effective metric space that represents HOI class prototypes, we propose SemantIc-Guided Multiple Attention Module (SIGMA-Module), which applies semantic information to highlight the discriminative regions in images. Since there are a noun and a verb in a description of HOI, a key challenge is how to coordinate their influence on attention. By doing so, we apply the idea of attention-erasing that focuses on the region of interest without being disturbed by previous attention guidance. In this way, the next semantic guidance is carried out on the features with attention-erasing. Therefore, the fusion of both attention parts outputs the highlighted visual regions, as illustrated in Fig. 3.

The specific process is as follows. First, a semantic-guided network f_ϕ is constructed to achieve semantic-guided attentive features:

$$\widetilde{s_i} = f_\phi(s_i, n_i, v_i), \tag{1}$$

where s_i is the i-th sample from the support set at episode e, n_i and v_i are noun and verb label semantic vectors of s_i respectively, and $\widetilde{s_i}$ is the final visual feature with multiple attention. It consists of 3 steps:

(1) We embed the category semantic information into the visual space as a highlight mask to guide visual features:

$$s_i^n = \sigma((1 + h(n_i))s_i), \tag{2}$$

where $h(\cdot)$ is a nonlinear network mapping semantic vectors to visual space, $\sigma(\cdot)$ is ReLu activate function, s_i^n is presentation features of s_i with noun-guided attention. In order to avoid influencing the next semantic-guided step, we subtract the feature with noun-guided attention over a threshold from the original image features.

(2) A step function is applied to preserve the high value of features as noun-guided attention since attention-erasing can

make networks focus on the non-overlapping region of interest. Particularly, we calculate the difference between initial visual features and noun-guided attention:

$$s'_i = s_i - \epsilon(\tau) \cdot s_i^n, \tag{3}$$

where s'_i is the output of attention-erasing operation, $\epsilon(\tau)$ is a step function. When τ is greater than a threshold, $\epsilon(\tau) = 1$, else $\epsilon(\tau) = 0$. Similarly, we obtain verb-guided attention by:

$$s_i^v = \sigma((1 + h(v_i))s'_i), \tag{4}$$

where v_i is verb label semantic of s_i, s_i^v is the feature with verb-guided attention.

(3) Noun and verb attention features are added up to obtain the integral visual features $\widetilde{s_i}$:

$$\widetilde{s_i} = s_i^n + s_i^v. \tag{5}$$

It should be noted that the samples in the query set are also processed by f_ϕ as that in support set. Since the semantic information is unavailable in query set, we train a semantic generation network to produce semantic vectors from images:

$$\widetilde{n}, \widetilde{v} = g(q), \tag{6}$$

where q is the sample from the query set, $g(\cdot)$ is semantic generation network that is a nonlinear neural network, $\widetilde{n}$ and $\widetilde{v}$ are generated noun and verb label semantic vectors of q. The output of the semantic generation network is forced to be close to the corresponding category semantic features, which is formulated as:

$$Loss_g = ||\widetilde{n} - n_c||_2^2 + ||\widetilde{v} - v_c||_2^2, \tag{7}$$

where n_c and v_c are respectively the true noun and verb label of q. The samples in query set are input into f_ϕ to get semantic-guided attentive features:

$$\widetilde{q} = f_\phi(q, \widetilde{n}, \widetilde{v}), \tag{8}$$

where $\widetilde{q}$ is the final image representations of sample q in query set. In this way, the class prototypes are guided discriminatively by the class semantic information.

Prototypes Shift Module

The features obtained from the SIGMA-Module represent more effective class prototypes. Ideally, a query sample is near to its class prototype, which is obtained by averaging samples in the support set (Snell, Swersky, and Zemel 2017). A more strict standard is that it will produce a small influence on original class prototypes if computing the query sample into prototypes, as shown in Fig. 4. Considering the complex HOI scene may generate an unstable prototype, we design a Prototypes Shift Module (PS-Module) under this strict standard, where there are different decision methods when training and testing. Specifically, we apply a super-parameter regulating shift distance in training to obtain a new prototype by fusing query samples with class prototypes. In testing, HOI recognition is performed by computing the direct distance to class prototypes, which achieves superior performance than training.

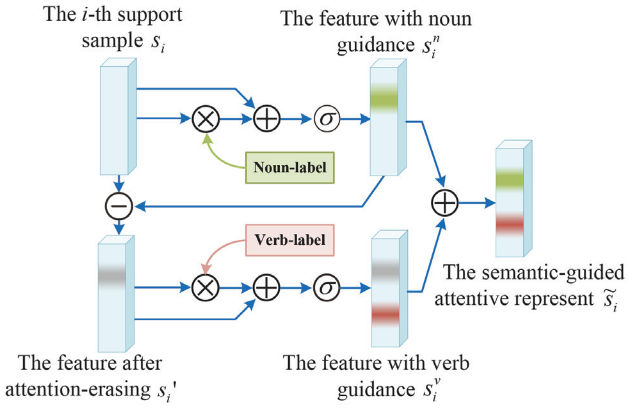

The i-th support sample s_i
The feature with noun guidance s_i^n
Noun-label
Verb-label
The semantic-guided attentive represent $\widetilde{s}_i$
The feature after attention-erasing s_i'
The feature with verb guidance s_i^v

Figure 3: Illustration of SIGMA-Module.

There are two types of prototypes in this module, i.e., original prototypes and prototypes shift. Concretely, the original prototypes p_c (of category c) are calculated as:

$$p_c = \frac{1}{N_S} \sum_{(s_i, y_i) \in S_e^c} f_\phi(s_i, n_i, v_i), \tag{9}$$

where N_S is the number of support set in class c and S_e^c is the support samples of class c in the episodic e. The prototypes shift $\widetilde{p_c}$ is calculated based on the trade-off between the query sample and original prototypes p_c:

$$\widetilde{p_c} = (1 - \alpha)p_c + \alpha f_\phi(q, \widetilde{n}, \widetilde{v}), \tag{10}$$

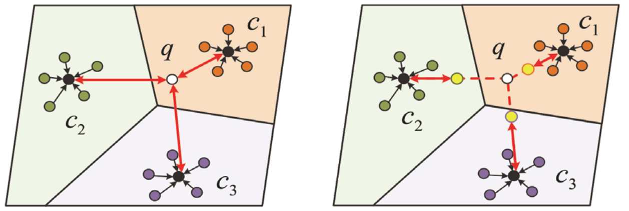
where α is a scaling factor to generate prototypes shift $\widetilde{p_c}$. The distance between $\widetilde{p_c}$ and p_c measures the perturbation of the query sample to each class prototype.

An image perhaps appears in the support set or the query set since images are randomly selected from a class. Therefore, the purpose of PS-Module is to make the prototype vector of the same class more consistent. By doing so, we introduce triplet loss (Schroff, Kalenichenko, and Philbin 2015) to strengthen the similarities in the same class and the differences among different classes:

$$Loss_{tr} = \sum_i^N [||f_\phi(q) - p_c^p||_2^2 - ||f_\phi(q) - p_c^n||_2^2], \tag{11}$$

where p_c^p and p_c^n are vectors of positive and negative prototypes. $Loss_{tr}$ is the triplet loss that helps to learn a good embedding where similar images are close and different images are far away.

A hard training strategy would help to learn a generalizable learner (Shrivastava, Gupta, and Girshick 2016). Therefore, constraining the distance between $\widetilde{p_c}$ and p_c is more effective than that between q and p_c in training. Accordingly,



(a) Prototypes are computed as the mean of embedded support instances for each class in Prototypical Networks.

(b) The PS-Module measures the bias of prototypes produced by the query sample.

Figure 4: Illustration for Prototypical Networks and our proposed PS-Module. Class prototypes c_i are computed as the mean of embedded support examples in category i, q is a sample in query set and red lines are different distances to be trained (Best viewed in color).

we apply a cross-entropy loss to meature it:

$$Loss_{ps} = d(\widetilde{p_c}, p_c) + log \sum_{N_C} exp(-d(\widetilde{p_c}, p_c)), \qquad (12)$$

where $d(\cdot)$ is the Euclidean distance, $Loss_{ps}$ is the loss function that limits the distance from prototypes shift to original prototypes.

When testing in the meta-test set, semantic information is generated by the semantic generation network for samples both in support sets and query sets. And we calculate a probability distribution for a query sample distance to the prototypes of support sets to accomplish the recognition task:

$$p_\phi(y = c | q \in Q_e^c) = \frac{exp(-d(f_\phi(q, \widetilde{n}, \widetilde{v}), p_c))}{\sum_k exp(-d(f_\phi(q, \widetilde{n}, \widetilde{v}), p_k))}. \qquad (13)$$

Experiments

Experiment Setup

We fix parameters in the pretrained CNN of ResNet-18 (He et al. 2016) to obtain visual features, and update all parameters in SIGMA-Module. We use Adam for the optimization, in which the initial learning rate is 0.00001. The parameter of regularizer term is 0.01. We set the threshold of step function $\tau = 1.5$ and the scaling factor of PS-Module $\alpha = 0.5$. Besides, we apply Word2vector (Mikolov et al. 2013) to extract the semantic embeddings for the category labels.

Datasets

We evaluate our method on the popular HICO (Chao et al. 2015) and TUHOI (Le, Uijlings, and Bernardi 2014) datasets. Original datasets are divided into a training set and a test set in shared label space, which can not satisfy the need for our experiments. We apply 60/20/20 training/testing/validation splits that are disjoint in noun labels for reorganizing the datasets. Details of both datasets are described below.

HICO-FS. HICO dataset (Chao et al. 2015) is a dataset for Humans Interacting with Common Objects, in which each interaction consists of a verb-noun pair. We reorganize this dataset by removing images with the label of "no interaction" and keeping the main labels for images. We call the modified dataset HICO for few-shot task as HICO-FS, which consists of 42,109 images with 80 nouns, 92 verbs, and 377 interactions. Considering ubiquitous similarities among actions, we take different nouns as different tasks. We divide HICO-FS into a meta-train set with 45 nouns and 25,968 images, a meta-test set with 20 nouns and 9,146 images, and a meta-val set with 15 nouns and 9,146 images, which are disjoint in noun labels.

TUHOI-FS. Trento Universal Human-Object Interaction dataset (Le, Uijlings, and Bernardi 2014) is dedicated to actions of human in images extracted from ImageNet. In order to establish a more natural and realistic dataset, the dataset annotates actions from images in stead of collecting images for some predefined human actions. Similar to that of HICO dataset, we reorganize TUHOI to be TUHOI-FS, which consists of 9802 images with 95 nouns, 66 verbs, and 194 interactions. We divide TUHOI-FS into a meta-train set with 50 nouns and 4871 images, a meta-test set with 25 nouns and 2570 images, and a meta-val set with 20 nouns and 2361 images. Each pair label in all meta-sets contains 6 samples at least and 487 samples at most to ensure that it can perform on both 5-way 1-shot and 5-way 5-shot tasks.

Comparison with State-of-the-Art Methods

The few-shot HOI recognition accuracies are computed by averaging 10 times over 600 randomly generated episodes. Five metric-based methods are chosen for comparison: Matching Networks (Vinyals et al. 2016), Prototypical Networks (Snell, Swersky, and Zemel 2017), Relation Networks (Sung et al. 2018), DN4 (Li et al. 2019a) and TPN (Liu et al. 2019), and two initialization-based method: MAML (Finn, Abbeel, and Levine 2017) and Reptile (Nichol, Achiam, and Schulman 2018). They all utilize ResNet-18 (He et al. 2016) as embedding networks.

From Table 2, we observe that our model achieves competitive performance for both tasks on HICO-FS dataset. Concretely, our method achieves accuracy of 38.16% in terms of 5-way 1-shot and 58.39% in terms of 5-way 5-shot, which outperforms the state-of-the-art approaches at least in 4.2% and 11.1%. It is worth noting that prior knowledge can be incorporated into metric-based approaches to improve their performance (Xing et al. 2019). To this end, our SGAP-Net utilizes semantic information to learn attentive class prototypes, which pay attention to actions and objects corresponding to HOI. It proves that semantic information can guide to learn a good metric space in few-shot HOI. Moreover, our SGAP-Net, which is formulated as a few-shot task in meta-learning framework, alleviates the challenges of instance imbalance and combinatorial explosion in HOI recognition.

Similar results are achieved on the TUHOI-FS dataset, as shown in Table 3. It demonstrates that our proposed SGAP-Net is capable of bringing about 3.4% and 13.3% performance gain respectively on 5-way 1-shot and 5-way 5-shot. However, the whole performance of all methods on TUHOI-FS is lower than that on HICO-FS. We suppose the reason

Table 2: Few-shot classification accuracy on test split on HICO-FS with ± 95% confidence intervals.

Methods	Type	5-way 1-shot	5-way 5-shot
Matching Networks (Vinyals et al. 2016)	Metric	32.14 ± 1.62%	44.87 ± 1.74%
Prototypical Networks (Snell, Swersky, and Zemel 2017)	Metric	32.56 ± 1.59%	42.49 ± 1.75%
Relation Networks (Sung et al. 2018)	Metric	33.20 ± 1.68%	46.15 ± 1.81%
DN4 (Li et al. 2019a)	Metric	33.07 ± 1.43%	46.19 ± 1.74%
TPN (Liu et al. 2019)	Metric	33.40 ± 1.55%	46.33 ± 1.86%
MAML (Finn, Abbeel, and Levine 2017)	Initialization	33.87 ± 1.74%	47.25 ± 1.84%
Reptile (Nichol, Achiam, and Schulman 2018)	Initialization	33.26 ± 1.77%	46.56 ± 1.85%
SGAP-Net (Ours)	Metric	**38.16 ± 1.65%**	**58.39 ± 1.82%**

Table 3: Few-shot classification accuracy on test split of TUHOI-FS with ± 95% confidence intervals.

Methods	Type	5-way 1-shot	5-way 5-shot
Matching Networks (Vinyals et al. 2016)	Metric	32.48 ± 1.58%	40.04 ± 1.70%
Prototypical Networks (Snell, Swersky, and Zemel 2017)	Metric	31.12 ± 1.55%	39.26 ± 1.71%
Relation Networks (Sung et al. 2018)	Metric	33.50 ± 1.68%	41.15 ± 1.75%
DN4 (Li et al. 2019a)	Metric	32.49 ± 1.43%	41.75 ± 1.77%
TPN (Liu et al. 2019)	Metric	32.95 ± 1.59%	41.73 ± 1.79%
MAML (Finn, Abbeel, and Levine 2017)	Initialization	33.78 ± 1.64%	43.67 ± 1.79%
Reptile (Nichol, Achiam, and Schulman 2018)	Initialization	32.39 ± 1.81%	41.65 ± 1.93%
SGAP-Net (Ours)	Metric	**37.27 ± 1.61%**	**57.05 ± 1.73%**

lies in the original distribution of data: the average samples of TUHOI-FS are much less than those of HICO-FS. Therefore, the few-shot HOI task is more difficult on TUHOI than that on HICO. Our work makes an attempt to apply semantic information to generate class prototypes for few-shot learning in HOI scene. And our results on both datasets demonstrate the effectiveness of our approach in complex human-centered scenarios.

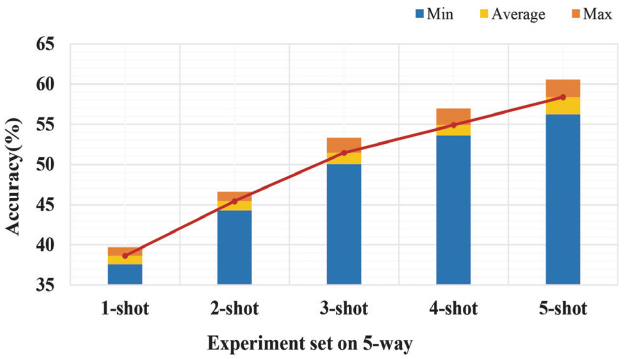


Figure 5: Visualization of performance with different experimental sets on HICO-FS.

We visualize the performance with different experimental sets on HICO-FS dataset to test the stability of our model, as shown in Fig. 5. We test SGAP-Net 10 times from 1-shot to 5-shot of 5-way. It can be observed that the bias between the maximum and the minimum on 5-way 1-shot is the smallest one. With the increase of shots, the fluctuations of models increase gradually. The average values are computed as the accuracy of the experiment sets (red line in Fig. 5). The most significant improvements on average value occur on 5-way 2-shot, in which it increases over 7% than that on 5-way 1-shot. It is due to the fact that more samples from classes provide strong support to the classifier.

We also examine experimental results on those categories shown in Fig.1, where categories in blue are in the training set and the others are in the test/val set. We take Prototypical Networks (Snell, Swersky, and Zemel 2017) as a baseline to provide the experimental evidence. Concretely, Prototypical Networks achieve the accuracy of 28.65% on 'hold-orange' and 39.76% on 'eat-orange' in terms of 5-way 1-shot. By contrast, our SGAP-Net achieves accuracy of 35.39% on 'hold-orange' and 72.97% on 'eat-orange' in terms of 5-way 1-shot, which outperforms Prototypical Networks 6.7% and 33.2%, respectively. For the task of 5-way 5 shot, our SGAP-Net achieves 9.56% and 26.1% performance improvement on 'hold-orange' and 'eat-orange'. Similar results are achieved on other categories, which brings significant performance gain on 5-way 1-shot and 5-way 5-shot.

Ablation Studies

We conduct ablation studies to evaluate the impacts of each component in our SGAP-Net in Table 4. We consider the following variants:

PN is Prototypical Networks (Snell, Swersky, and Zemel 2017) as the baseline for SGAP-Net.

SGAP (w Verb) is the PN that only includes the verb labels.

SGAP (w Noun) is the PN that only includes the noun labels.

SGAP (w/o AE) utilizes information of verb and noun by adding linearly without attention-erasing.

SGAP (w/o PS) replaces the PS-Module by the desicion method in PN.

Table 4: Ablation studies of SGAP-Net on HICO-FS.

Methods	5-way 1-shot	5-way 5-shot
PN	$32.56 \pm 1.59\%$	$42.49 \pm 1.75\%$
SGAP (w Verb)	$35.86 \pm 1.63\%$	$54.49 \pm 1.86\%$
SGAP (w Noun)	$35.56 \pm 1.66\%$	$54.68 \pm 1.87\%$
SGAP (w/o AE)	$32.03 \pm 1.56\%$	$49.06 \pm 1.84\%$
SGAP (w/o PS)	$37.48 \pm 1.67\%$	$56.79 \pm 1.90\%$
SGAP-Net(ours)	$\mathbf{38.16 \pm 1.65\%}$	$\mathbf{58.39 \pm 1.87\%}$

We can observe that applying a single type of semantic information improves the result over 3% compared with Prototypical Networks, as shown in Table 4. It is proved that semantic-guided attention mechanism is effective for few-shot HOI recognition. If both noun and verb information are included in the model and are combined linearly, we observe that it performs even 0.4% worse than Prototypical Networks on the 5-way 1-shot task. This is a reasonable phenomenon since stacking attention representations will entangle with each other. Then we utilize verb and noun semantic embedding based on attention-erasing and obtain improvement with almost 6% than Prototypical Networks. Meanwhile, we find results are almost unaffected when exchanging the order of two types of semantic information, which demonstrates that the method of attention-erasing works effectively on twice guidance.

PS-Module on miniImageNet

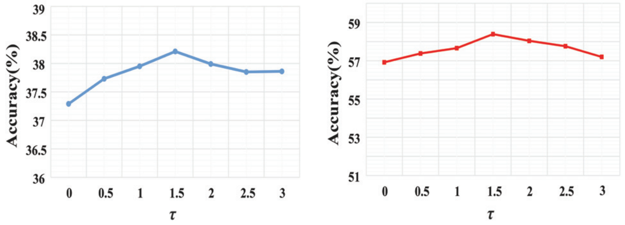
To prove the effectiveness of the PS-Module on the general few-shot learning task, we evaluate our PS-Module on miniImageNet (Vinyals et al. 2016), which is a popular dataset for few-shot learning. The performance is summarized in Table 5. They both utilize 4 layers of CNN to extract visual features. Our PS-Module achieves the improvement of 0.8% on the 5-way 1-shot task and 1.1% on the 5-way 5-shot task on miniImageNet dataset. It demonstrates that our proposed PS-Module is also capable of improving the performance of conventional few-shot classification. Simultaneously, we find PS-Module reduces the variance of accuracies by employing a strict decision method, which strengthens the stability of our model to perform in the real scene.

Table 5: The effectiveness of PS-Module on miniImageNet.

Methods	5-way 1-shot	5-way 5-shot
Prototypical Net	$44.32 \pm 0.86\%$	$63.73 \pm 0.72\%$
PS-Module	$\mathbf{45.14 \pm 0.76\%}$	$\mathbf{64.85 \pm 0.68\%}$

Quantitative analysis

Threshold τ in step function. An experiment is implemented to further evaluate the influences of the threshold τ of the step function in Eq.(3). We fix the scaling parameter at the default value ($\alpha = 0.5$) and vary the threshold of step function τ from 0 to 3. The corresponding experimental results are presented in Fig. 6. This step function is designed to preserve highlight information of features. From the results, we find that the performance of SGAP-Net with different thresholds of step function shows the same trend on both experiment sets, which achieves the apex when $\tau = 1.5$. Besides, when there is no parameter to control the intensity of attention-erasing ($\tau = 0$), the performance of the models is the worst. With τ increasing over 1.5, the results also show a decline. We consider that inadequate attention-erasing may cause bias of feature representation.

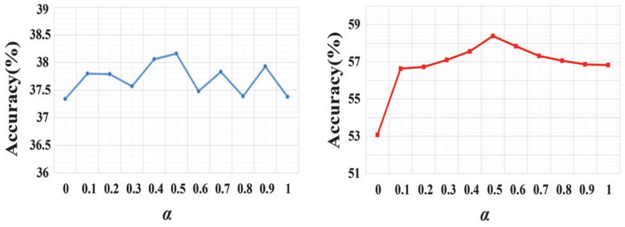


(a) The few-shot HOI recognition results on 5-way 1-shot.

(b) The few-shot HOI recognition results on 5-way 5-shot.

Figure 6: Impact of varied thresholds τ on HICO-FS dataset.

Scaling parameter α. We further investigate the influence of parameter α in Eq. (10), which controls the scaling parameter of measuring the distance prototypes shift, and the results are shown in Fig. 7. It can be observed that the performance is improved continuously with the increase of α until the performance reaches their peak when α = 0.5 on the task of 5-way 5-shot and then start to decrease gradually. The performance on the task of 5-way 1-shot is different from that on 5-way 5-shot. It is due to that the only support sample provided by the task of 5-way 1-shot can not generate a stable class prototype in metric space. Despite obvious fluctuations on the task of 5-way 1-shot, our model also achieves the best performance at α = 0.5. Through the observation, we can conclude that incorporating an appropriate ratio of query samples into prototypes can steadily improve the performance of few-shot HOI recognition.



(a) The few-shot HOI recognition results on 5-way 1-shot.

(b) The few-shot HOI recognition results on 5-way 5-shot.

Figure 7: Impact of varied scaling parameter α on HICO-FS dataset.

Conclusion

This paper formulated HOI recognition as a few-shot learning task into a meta-learning framework to alleviate instance imbalance and combinatorial explosion challenges. In this work, we have proposed SGAP-Net that learns a metric space where HOI recognition can be performed by computing distances to prototype representations of each class.

The SIGMA-Module in it generates discriminative prototypes via erasing and stacking attention from the class semantic of verb and noun labels. And the PS-Module in it is designed by different decision methods in training and testing, which achieves performance gain by a super-parameter regulating shift distance. In addition, Two human-object interaction datasets, HICO-FS and TUHOI-FS, are released to few-shot HOI. Extensive experiments have demonstrated that our proposed SGAP-Net is superior to the state-of-the-art approaches.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 61771329 and 61632018, and the Natural Science Foundation of Tianjin under Grant 19JCYBJC16000.

References

Annadani, Y., and Biswas, S. 2018. Preserving semantic relations for zero-shot learning. In *Conference on Computer Vision and Pattern Recognition*, 7603–7612.

Chao, Y.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *International Conference on Computer Vision*, 1017–1025.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135.

Fu, J.; Liu, J.; Tian, H.; Fang, Z.; and Lu, H. 2018. Dual attention network for scene segmentation. In *Conference on Computer Vision and Pattern Recognition*, 3146–3154.

Girdhar, R., and Ramanan, D. 2017. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, 34–45.

Gkioxari, G.; Girshick, R. B.; Dollar, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *Conference on Computer Vision and Pattern Recognition*, 8359–8367.

Gkioxari, G.; Girshick, R. B.; and Malik, J. 2015. Contextual action recognition with r*cnn. In *International Conference on Computer Vision*, 1080–1088.

Gupta, A., and Davis, L. S. 2007. Objects in action: An approach for combining action understanding and object perception. In *Conference on Computer Vision and Pattern Recognition*, 1–8.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. B. 2017. Mask R-CNN. In *International Conference on Computer Vision*, 2980–2988.

Ji, Z.; Xiong, K.; Pang, Y.; and Li, X. 2019. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology* 1:1–10.

Kato, K.; Li, Y.; and Gupta, A. 2018. Compositional learning for human-object interaction. In *European Conference on Computer Vision*, 234–251.

Le, D.; Uijlings, J. R. R.; and Bernardi, R. 2014. TUHOI: Trento universal human object interaction dataset. In *Workshop on Vision and Language*, 17–24.

Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; and Luo, J. 2019a. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Conference on Computer Vision and Pattern Recognition*, 7260–7268.

Li, Y.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.; Wang, Y.; and Lu, C. 2019b. Transferable interactiveness knowledge for human-object interaction detection. In *Conference on Computer Vision and Pattern Recognition*, 3585–3594.

Lin, B.; Kan, L.; Pei, J.; and Shuai, J. 2016. Main objects interaction activity recognition in real images. *Neural Computing & Applications* 27(2):335–348.

Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S. J.; and Yang, Y. 2019. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations*, 1–14.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv: 1803.02999*.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137–1149.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition*, 815–823.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *Conference on Computer Vision and Pattern Recognition*, 761–769.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Vinyals, O.; Blundell, C.; Lillicrap, T. P.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 3637–3645.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Conference on Computer Vision and Pattern Recognition*, 6450–6458.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Conference on Computer Vision and Pattern Recognition*, 6488–6496.

Xing, C.; Rostamzadeh, N.; Oreshkin, B. N.; and Pinheiro, P. H. O. 2019. Adaptive cross-modal few-shot learning. *arXiv preprint arXiv: 1902.07104*.

Yu, Y.; Ji, Z.; Fu, Y.; Guo, J.; and Zhang, Z. 2018. Stacked semantic-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems*, 4321–4329.

Zhuang, B.; Wu, Q.; Reid, I. D.; Shen, C.; and Den Hengel, A. V. 2018. HCVRD: a benchmark for large-scale human-centered visual relationship detection. In *the 32nd AAAI Conference on Artificial Intelligence*, 7631–7638.