

Shuffle-Then-Assemble: Learning Object-Agnostic Visual Relationship Features

Xu Yang, Hanwang Zhang, Jianfei Cai

School of Computer Science and Engineering,
Nanyang Technological University,
s170018@e.ntu.edu.sg, {hanwangzhang, asjfcai}@ntu.edu.sg

Abstract. Due to the fact that it is prohibitively expensive to completely annotate visual relationships, *i.e.*, the (obj1, rel, obj2) triplets, relationship models are inevitably biased to object classes of limited pairwise patterns, leading to poor generalization to rare or unseen object combinations. Therefore, we are interested in learning object-agnostic visual features for more generalizable relationship models. By “agnostic”, we mean that the feature is less likely biased to the classes of paired objects. To alleviate the bias, we propose a novel **Shuffle-Then-Assemble** pre-training strategy. First, we discard all the triplet relationship annotations in an image, leaving two unpaired object domains without obj1-obj2 alignment. Then, our feature learning is to recover possible obj1-obj2 pairs. In particular, we design a cycle of residual transformations between the two domains, to capture shared but not object-specific visual patterns. Extensive experiments on two visual relationship benchmarks show that by using our pre-trained features, naive relationship models can be consistently improved and even outperform other state-of-the-art relationship models. Code has been made available at: <https://github.com/yangxuntu/vrd>.

1 Introduction

Thanks to the maturity of mid-level vision solutions such as object classification and detection [19,41,15], we are more ambitious to pursue higher-level vision-language tasks such as image captioning [13,14,5,31], visual Q&A [22,27,18], and visual chatbot [7]. Unfortunately, we gradually realize that many of the state-of-the-art systems merely capture training set bias while not the underlying reasoning [49,22,65]. Recently, a promising way is to use visual compositions such as scene graph [23,53] and relationship context [21,62] for explainable visual reasoning. Therefore, visual relationship detection (VRD) [60,61,28,57] — the task of predicting elementary triplets such as “person ride bike” and “car park on road” in an image — is becoming an indispensable building block connecting vision with language.

Despite the relatively preliminary stage of VRD compared to object detection, a major challenge of VRD is the high cost of annotating the (obj1, rel, obj2) triplets as shown in Fig. 1 (a). Unlike labeling objects in images, labeling visual

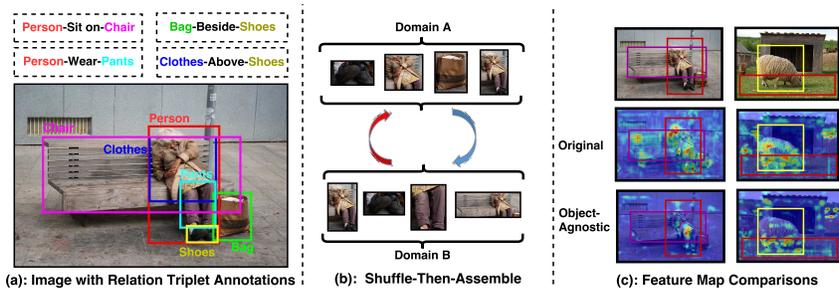


Fig. 1: (a) The triplet annotations of visual relationships in an image. (b) The key idea of the proposed **Shuffle-Then-Assemble** strategy is to discard the paired annotation of any relationship and leave two unpaired object domains. (c) Comparisons between the original feature maps obtained from base CNN (middle) and the object-agnostic ones (bottom) obtained by our pre-training (averaged over all the channels). We can see that our feature maps focus more on the overlapped regions of objects.

relationships is prohibitively expensive as it requires combinatorial checks of the three entries. Therefore, the relationships in existing VRD datasets [32,26] are long-tailed, and the resultant relationship models are inevitably biased to the dominant obj1-obj2 combinations. For example, as reported in pioneering works [60,32], the recognition rate of unseen triplet compositions is significantly lower than the seen ones. This deficiency clearly limits the VRD potential in compositional reasoning. Though it can be alleviated by exploiting external knowledge such as language priors [32] and large-scale weak supervision [61], we still lack a principled solution in the visual modeling perspective.

Unsupervised feature learning (or pre-training) is arguably the most popular remedy for training deep models with small data [64,36,10,39,11,47]. Therefore, we are inspired to learn object-agnostic convolutional feature maps that are less likely biased to certain obj1-obj2 combinations. Such features should be highly responsive to object parts¹ involved in a relationship. A plausible way is to append additional conv-layers to the original base CNN (*e.g.*, VGG16 [44] or ResNet-150 [19]) to remove the object-sensitive responses inherited from image classification pre-training dataset (*e.g.*, ImageNet [8]). For example, as shown in Fig. 1 (c), compared with the base CNN’s feature map, the object-agnostic one ignores object patterns but focuses on the shared patterns of interacted objects. Therefore, we raise a question: how to learn the object-agnostic feature maps without additional relationship labeling cost?

In this paper, we propose a novel **Shuffle-Then-Assemble** feature learning strategy. As shown in Fig. 1 (b), “shuffle” is to discard the original one-to-one paired object alignments, and thus no explicit obj1-obj2 class information is used; “assemble” is to pose the relationship modeling into an unsupervised

¹ The parts can be at the pixel-level as well as the receptive field-level.

pair recover problem by transferring Region-of-Interest (ROI) features between the two unpaired domains. Our intuitive motivation is two-fold: 1) if the ROI features extracted from the resultant feature maps still encode object-specific information, features are not likely to be transferred between the two domains of heterogeneous objects; 2) the unsupervised fashion encourages the exploration of many more possible relationships which are usually missing in annotation. As shown in Fig. 1 (a), some simple spatial relationships such as “chair beside bag” are missing, and equivalent relationships are usually ignored, *i.e.*, “chair under person” is missing as “person sit on chair” is labeled. Inspired by the recent advances in unsupervised domain transfer methods [66,24,20,56], we design a cycle of transformations to establish the transfer between the two domains: either transfer direction maps an RoI from domain A (or B) to B (or A), and then an adversarial network is used to confuse the mapping with RoIs in B (or A). In particular, we use a residual structure for the transformation network, where the identity mapping encourages the feature map to capture shared but not object-specific visual patterns and the residual allows feature transformation.

We demonstrate the effectiveness of the proposed **Shuffle-Then-Assemble** strategy on two benchmarks: VRD [32] and VG [26]. We observe consistent improvement of using our pre-trained features against various ablative baselines and other state-of-the-art methods. For example, compared to feature maps without pre-training, we can boost the Recall@100 of supervised, weakly supervised, and zero-shot relationship prediction by absolute 4.74%, 4.42%, 4.04%, respectively on VRD, and 4.41%, 4.2%, 5.81%, respectively on VG.

2 Related Work

Visual Relationships. Modeling the object interactions such as verbs [16,3], actions [17,40,54], and visual phrases [55,1,43,9] has been extensively studied in literature. In particular, our relationship model used in this paper follows the recent progress on modeling generic visual relationships, *i.e.*, the (obj1, rel, obj2) triplets detected in images [32,60]. State-of-the-art relationship models fall into two lines of efforts: 1) message passing between the two object features [57,28,52], and 2) exploitation of subject-object statistics such as language priors [32,29,67] and dataset bias [59,63,6]. However, they are still limited in the inherent issue of insufficient training triplets due to combinatorial annotation complexity, leading the resultant relationship model to be brittle to rare or unseen compositions. Though weakly-supervised methods [61,38,50] can reduce the labeling cost, its performance is still far from practical use compared to supervised models. Unlike previous methods, in this paper, we propose to resolve this challenge in pairwise modeling of relationship, that is, given two regions, we want to improve the predicate classification without additional object information and extra supervision. We believe that the improvement can boost most of the above relationship models by replacing their pairwise modeling counterparts with our method.

Unsupervised Feature Learning. By exploiting large-scale unlabeled data, unsupervised feature learning methods [2] learn more generalizable intermediate

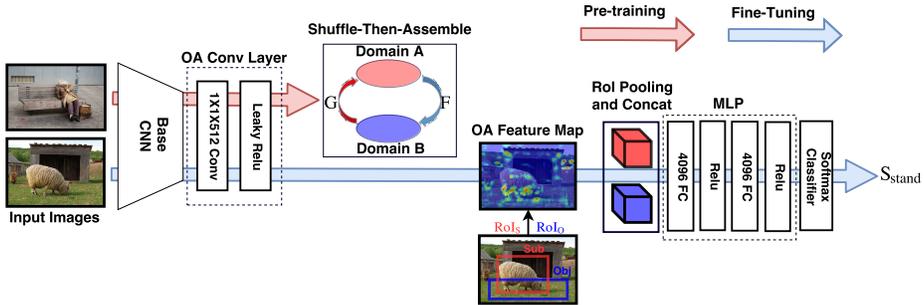


Fig. 2: The overview of the proposed **Shuffle-Then-Assemble** pre-training strategy (red arrow) and relationship detection model (blue arrow). The goal at the pre-training stage is to learn the Object-Agnostic (OA) conv-layers using **Shuffle-Then-Assemble** objective. Then, the traditional supervised training for the relation model can be considered as the fine-tuning stage using the desired OA feature map.

data representation for solving some other machine learning tasks. Our motivation for visual relationship feature learning follows the common practice: feature transfer in today’s computer vision [58], which fine-tunes a base network which has been pre-trained on other datasets and tasks. Different from the popular auto-encoder fashion [64,11], our strategy is more similar to the recent works on self-supervised training, where the learning objective is to discover the inherent data compositions such as predicting the context of image patches [10,36,37,33,34,45]. In particular, we propose to discover the alignment of RoI pairs and pose this discovery into the task of unsupervised domain transfer using adversarial learning [66,24,20,56]. Inspired by them, we use a cycle of transformations to remove the trivial alignment caused by mode collapse and thus build non-trivial connections between the paired RoIs.

3 Method

Fig. 2 illustrates the overview of using **Shuffle-Then-Assemble** to enhance the relationship model. The goal of the feature learning process is to pre-train the Object-Agnostic (OA) conv-layers, which result in the desired OA feature map for better relationship modeling. We will first introduce the widely-used relationship modeling framework and its limitations, and then detail how to use the proposed feature learning method to overcome them.

3.1 Visual Relationship Model

The input of the visual relationship model is an image with a pair of object bounding boxes, and the output is an “obj1-rel-obj2” triplet, where “obj1” and

“obj2” are the object classes of the two bounding boxes, and “rel” is the relationship class. In this paper, we adopt the common practice as in [32,60] that we do not directly model the triplet composition as a whole [43,6], which requires $\mathcal{O}(C^2R)$ complexity for C object and R relationship classes; instead, we model objects and relationships separately to reduce the complexity down to $\mathcal{O}(C+R)$. Therefore, without loss of generality, we refer to a relationship model as an R -way classifier.

Suppose \mathbf{x}_i and \mathbf{x}_j are the RoI features of any pair of object bounding boxes (i, j) (e.g., the red and blue cubes in Fig. 2 by RoI pooling [12]), the r -th relationship score is obtained by a softmax classifier whose input is a simple concatenation of the two features:

$$S(i, j, r; \theta) = \frac{\exp(\mathbf{w}_r^T \text{MLP}([\mathbf{x}_i, \mathbf{x}_j]))}{\sum_{t=1}^R \exp(\mathbf{w}_t^T \text{MLP}([\mathbf{x}_i, \mathbf{x}_j]))}, \quad (1)$$

where $\mathbf{w}_t \in \theta$ is the parameter of the classifier and the configuration of $\text{MLP}(\cdot)$ is detailed in Fig. 2. Note that although Eq. (1) is a naive model and there are fruitful ways of combining \mathbf{x}_i and \mathbf{x}_j in the literature, such as appending independent MLPs for each RoI [60], the union RoI [28], and even the fusion with textual features [21], our feature learning can be seamlessly incorporated into any of them. We will leave the evaluations of applying these tweaks for future work.

The relationship model can be trained by minimizing the cross-entropy loss of Eq. (1), summing over all the relationship pairs. However, due to the limited annotation of the relationship triplets, relationship models trained on these extremely long-tailed annotations are inevitably biased to the dominant object classes. One may wonder why it is object-biased as Eq. (1) does not use any object class information at all? The reason resides in the base CNN feature map. Almost all state-of-the-art visual recognition systems deploy the base CNN [46,44,19] pre-trained on ImageNet [8] or ImageNet+MSCOCO [30], where the training task is object recognition. Therefore, the resultant feature map for extracting RoI will naturally favor the sensitivity to object classes — each RoI feature encodes the discriminative information of the object inside the RoI (cf. the original feature map of Fig 2), and leads the parameters in Eq. (1) over-fitted to specific object patterns. For example, if most of the triplets of “stand on” is “person stand on street”, then the “stand on” classifier will mistakenly consider the joint pattern “person” and “street” into “stand on”, and fails in cases of “person stand on chair” or “dog stand on street”.

3.2 Shuffle-Then-Assemble Feature Learning

To alleviate the bias, we detail our proposed **Shuffle-then-Assemble** strategy to pre-train the Object-Agnostic (OA) conv-layers for obtaining the OA feature map. As discussed above, the bias is mainly due to the dominant object pairs in training data, therefore, our key idea is to discard the original one-to-one

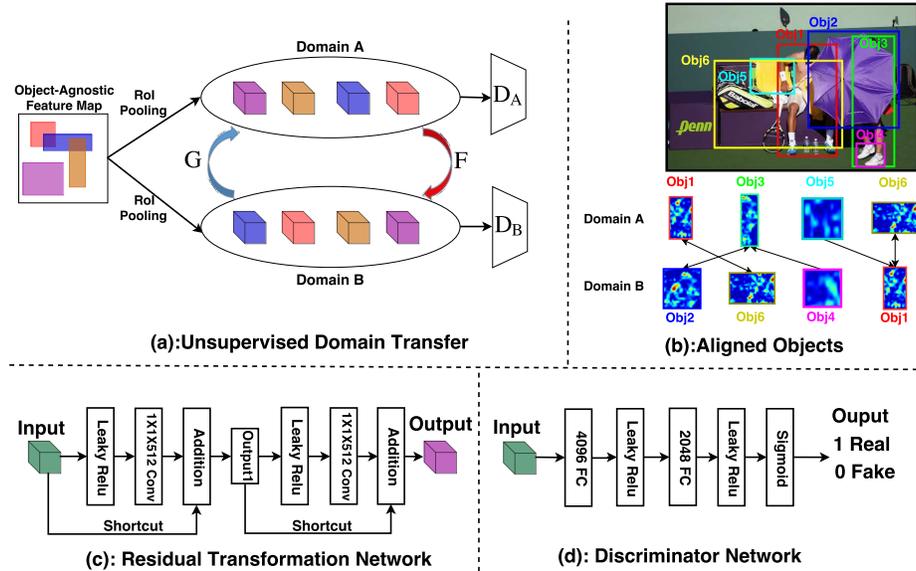


Fig 3: (a) The overview of unsupervised domain transfer for **Shuffle-Then-Assemble**. It contains a cycle of transformations $F: A \mapsto B$ and $G: B \mapsto A$, and a pair of discriminators D_A and D_B to measure the quality of the transfer. (b) Qualitative transfer results. The directed arrow indicates the nearest-neighbor RoI in the target domain to the RoI from the source domain. (c) The residual architecture of the transformation network. (d) The architecture of the discriminator.

pairwise annotations, *i.e.*, “shuffle”, leaving two unaligned domains of RoIs for “obj1” and “obj2”, and then we attempt to recover the one-to-one alignment, *i.e.*, “assemble”, by unsupervised domain transfer. Note that this pre-training strategy does not require additional cost of supervision. As shown in Fig. 3 (b), we manage to align potential relationships without any one-to-one supervision, *e.g.*, obj1 may relate to obj6 with respect to “sit” and obj3 may relate to obj2 with respect to “hold”.

The unsupervised domain transfer method used in **Shuffle-Then-Assemble** follows recent progress on adversarial domain transfer [66,24,20,56,4]. Noteworthy, the motivation of using adversarial domain transfer emphasizes more on the unsupervised alignments but NOT the feature transfer as in traditional domain transfer applications such as [48], where the domain transfer is used to close the gap between conveniently available synthetic data and real data. Here, our idea is more similar to [39] which discovers alignments between images that are very visually different such as “spotted bags” and “spotted shoes”, or “frontal faces” and “frontal cars”.

As illustrated in Fig 3 (a), we want to guide the pre-training of the OA conv-layers by learning mapping functions between domain A and B , where each of

them consists of RoI features, $\mathbf{a} \in A$ and $\mathbf{b} \in B$, extracted from the tentative OA feature map. For the purpose of domain transfer, we have a cycle of two mappings: $F: A \mapsto B$ and $G: B \mapsto A$, to discover the underlying relationship between A and B . Recall that there is one-to-one supervision between the two domains, we adopt the adversarial objective \mathcal{L}_{adv} such that the mapped features $\{F(\mathbf{a})\}$ and $\{G(\mathbf{b})\}$ are indistinguishable from B and A , respectively; in particular, the indistinguishability is measured by two discriminators D_A and D_B :

$$\mathcal{L}_{adv}(A, B; \phi, F, G, D_A, D_B) = \underbrace{\mathbb{E}_{\mathbf{a}}[\log D_A(\mathbf{a})] + \mathbb{E}_{\mathbf{b}}[\log D_B(\mathbf{b})]}_{\text{maximize by } D_A \text{ and } D_B} + \underbrace{\mathbb{E}_{\mathbf{b}}[\log(1 - D_A(G(\mathbf{b})))] + \mathbb{E}_{\mathbf{a}}[\log(1 - D_B(F(\mathbf{a})))]}_{\text{minimize by } F \text{ and } G}, \quad (2)$$

where ϕ is the OA conv-layers that generate A and B , D_A is a binary classifier that tries to classify $D_A(\mathbf{a}) \mapsto 1$ and $D_A(F(\mathbf{b})) \mapsto 0$, and D_B is defined similarly. In this adversarial way, we will eventually obtain F and G that discover the hidden alignment between the two domains, *i.e.*, indistinguishable by the discriminators.

To encourage more explorations of the potential relationship alignments between the RoIs in the two domains, *e.g.*, avoid from mapping many RoIs in A to only one RoI in B with respect to a trivial spatial relationship such as “on” and “by”, we further impose the “cycle-consistent” loss to be minimized by G and F :

$$\mathcal{L}_{cycle}(A, B; \phi, F, G) = \mathbb{E}_{\mathbf{a}}[\|\mathbf{a} - G(\mathbf{b})\|_1] + \mathbb{E}_{\mathbf{b}}[\|\mathbf{b} - F(\mathbf{a})\|_1]. \quad (3)$$

The loss penalizes two different RoIs, *e.g.*, \mathbf{a} and \mathbf{a}' , mapped to the same RoI \mathbf{b} as it is hard to satisfy $\mathbf{a} \approx G(\mathbf{b})$ and $\mathbf{a}' \approx G(\mathbf{b})$ simultaneously.

By putting Eq. (2) and Eq. (3) together, the full objective for pre-training the OA conv-layers is:

$$\phi^* = \arg \min_{\phi} \min_{F, G} \max_{D_A, D_B} \mathcal{L}_{adv}(A, B; \phi, F, G, D_A, D_B) + \lambda \mathcal{L}_{cycle}(A, B; \phi, F, G), \quad (4)$$

where $\lambda > 0$ is a trade-off hyper-parameter. Then, we can use ϕ^* to obtain \mathbf{x}_i and \mathbf{x}_j , and fine-tune a better relationship model θ using existing triplet supervision as in Eq. (1). Next, we will introduce the proposed implementation of F and G .

3.3 Implementation Details

Network Architecture. For base CNN, we adopt Faster RCNN (VGG16) [42], which takes short width to be 600 and outputs the original $1/16 \times 1/16 \times 512$ feature map. As shown in Fig 2, our OA conv-layer has 1 filter of the size 1×1 , stride 1, followed by a Leaky Relu [51]. The transformation network is detailed in Fig 3 (c). Each transformation contains two blocks of residual network. The

motivation of applying the residual structure is two-fold. 1) The shortcut encourages to find shared regions of two RoIs, since the shared RoI features will pass directly via the shortcut. This makes the optimization not only more light-weighted, but also easier to find the intrinsic inter-related visual patterns as relationships. 2) If any object-specific information is still encoded in the RoI feature, the shortcut will make it harder to achieve the final domain transfer as domain A and B usually contain diverse objects. The discriminator network is detailed in Fig 3 (d), which is composed by two fully-connected layers followed by Leaky Relu. It takes a 50,176-d (two $7 \times 7 \times 512$ RoI feature) vectorized RoI feature as input and outputs a sigmoidal scalar between 0 and 1.

Training. At the feature pre-training stage, to collect sufficient RoIs in each domain, we augment the number of original bounding boxes by additional ones with IoU larger than 0.7, extracted by using the Region Proposal Network [42]. For each original bounding boxes, 10 RoIs are sampled. To stabilize the adversarial training in Eq. (4), we adopt three practices: 1) We apply least-square GAN [35] to replace the negative log likelihood by a least square loss. 2) The optimizer for training D_A and D_B is set to SGD and the optimizer for G , F and ϕ is set to Adam [25]. The initial learning rate is set to $1e-4$ for both optimizers. 2) D_A and D_B are trained three times more compared with G , F and ϕ . The trade-off λ in Eq. (4) is set to 10. Every mini-batch is one image with randomly selected 128 triplets. The epochs for training these networks are set to 20 on VRD dataset and set to 5 on VG dataset.

At the fine-tune stage for training relationship classifier, the short width of image is still set to 600. Every mini-batch is one image with 128 randomly selected triplets. The optimizer is Adam with initial learning rate set to $1e-5$ in all the experiments. The epochs are set to 50 and 30 on VRD dataset and VG dataset, respectively.

4 Experiments

We evaluated our **Shuffle-Then-Assemble** method by performing visual relationship prediction on two benchmark datasets. We conducted experiments under extensive settings: supervised, weakly-supervised, and zero-shot, each of which has various ablative baselines and state-of-the-art methods. We also visualized qualitative object-agnostic features maps compared against others.

4.1 Datasets and Metrics

We used two publicly available datasets: VRD (Visual Relationships Dataset[32]) and VG (Visual Genome V1.2 dataset [26]).

VRD dataset. It contains 5,000 images with 100 object categories and 70 relationships. In total, VRD contains 37,993 relationship triplet annotations with 6,672 unique triplets and 24.25 relationship per object category. We followed the same train/test split as in [32], *i.e.*, 4,000 training images and 1,000 test images, where 1,877 triplets are only in the test set for zero-shot evaluations.

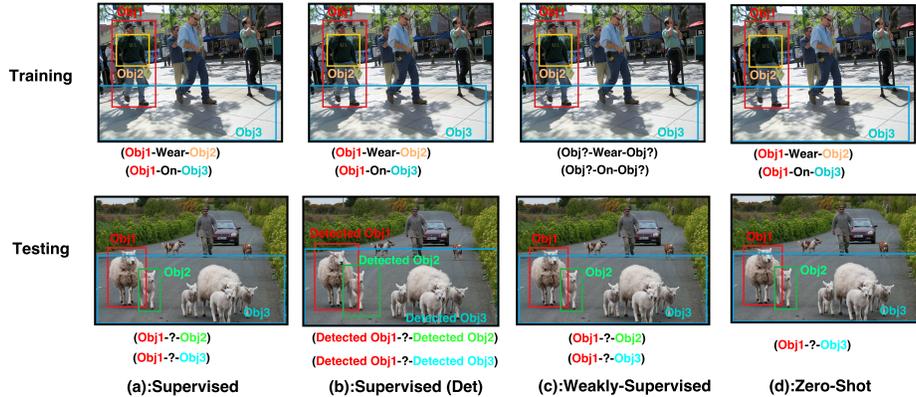


Fig. 4: We evaluate relationship prediction task using four different experiment settings: supervised, supervised (Det), weakly-supervised and zero-shot. “?” denotes the relationship to be predicted. It is noteworthy that the object category is not known under all the experiment settings, and we only use visual features to predict the relationship between object pairs.

VG dataset. We used the pruned version provided by Zhang[60] since the original one is very noisy. As a result, VG contains 99,658 images with 200 object categories and 100 predicates, 1,174,692 relation annotations with 19,237 unique relations and 57 predicates per object category. We followed the same 73,801/25,857 train/test split. And this dataset contains 2,098 relationships which never occur in the training set, which can be used for zero-shot evaluations.

Metrics. As conventions [32,60], we used Recall@50 (**R@50**) and Recall@100 (**R@100**) as evaluation metrics. R@K computes the fraction of times a true relationship is predicted in the top K confident relation predictions in an image.

4.2 Settings

In our experiments, we only focused on the relationship prediction task, *i.e.*, classifying any two object regions into relationship classes. The reasons are twofold. First, relationship prediction plays the core role in relationship detection, a more comprehensive task that also needs to detect the two objects. Second, we can exclude the influence of object detection performance, as the improvement of object detection can improve the relationship detection scores [60]. To offer a testbed for application domains of relationship prediction, we designed the following 4 settings according to different pairwise modeling fashions:

Supervised. This setting is the standard supervised relationship prediction. As shown in Fig 4 (a), for training, all the objects are provided with ground truth boxes and the relationship between objects are given; for testing, a pair of objects with bounding boxes are given and their relationship is to be predicted.

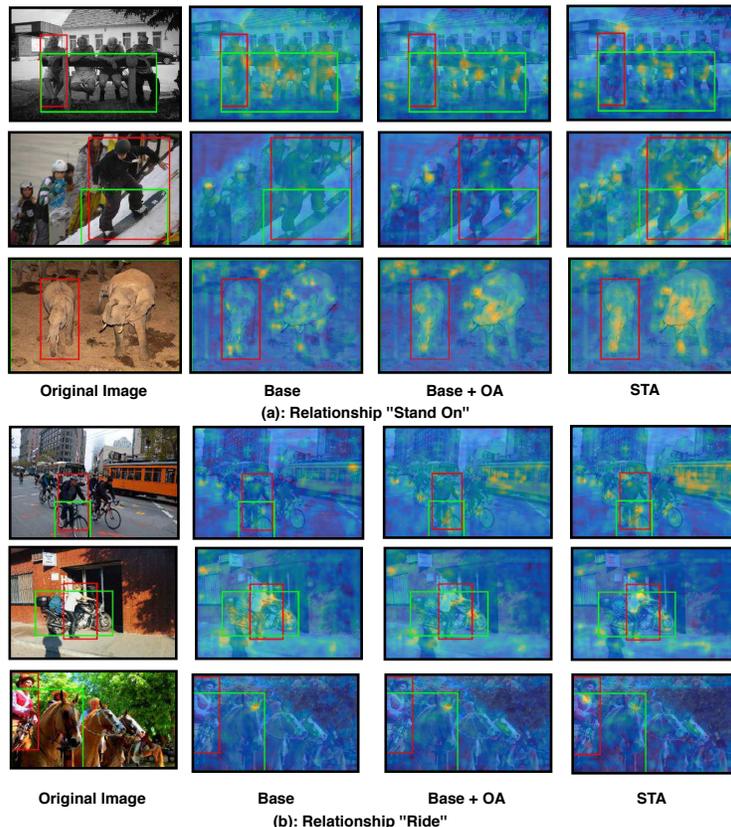


Fig. 5: Qualitative feature maps of two relationships on VRD dataset. For each one, three kinds of feature maps are visualized by averaging over the 512 channels. We can see that by using the proposed **Shuffle-Then-Assemble** (STA), the RoI features are less likely biased to the objects and more focused on the regions of interaction of the two objects. Moreover, the observation is consistent with diverse relationship appearances.

Supervised (Det). The above setting assumes a perfect object bounding box detector at testing. However, as shown in Fig 4 (b), a more practical setting is to use detected object bounding boxes using off-the-shelf object detectors. We used Faster RCNN [42] to detect around 100 objects in an image.

Weakly-Supervised. Compared to Supervised setting, we discard the one-to-one paired object annotation with respect to a relationship. As shown in Fig 4 (c), at training, given objects with boxes, we do not know which object relates to which one. Therefore, we used an average-pooling image-level relationship loss:

$$\mathcal{L}_{weak} = - \sum_{i=1}^N \sum_{j=1}^N \sum_{r=1}^R [y_{ijr} \log S(i, j, r) + (1 - y_{ijr}) \log(1 - S(i, j, r))]; \quad (5)$$

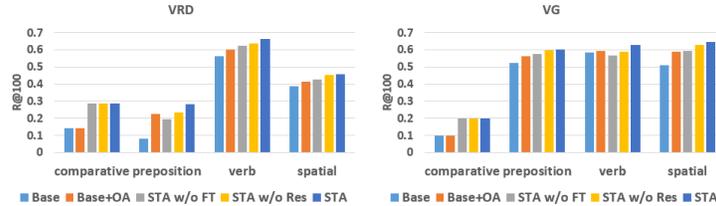


Fig. 6: Performances (R@100%) of relation classification of the four relation types using the different methods in the supervised setting.

where N is the number of objects, y_{ijr} is 1 if the object pair (i, j) has the r -th relationship, and $S(i, j, r)$ is the relationship score in Eq. (1). Note that the testing stage of this setting is the same as that of Supervised setting.

Zero-Shot. This setting is the same as Supervised setting except that at testing we want to predict object pairs whose triplet combination is unseen during training. As shown in Fig 4 (d), though object sheep, road, and relationship on are individually seen at training, but the composition “sheep on road” is novel to test.

Comparing Methods. We compared the proposed **Shuffle-Then-Assemble (STA)** pre-training strategy with the following ablative baselines:

Base. We directly use ROI features which extracted from the base CNN for relationship prediction task.

Base+OA. We do not pre-train OA conv-layers ϕ (in Eq. (2)) by **Shuffle-Then-Assemble** strategy and directly fine-tune ϕ and $MLP(\cdot)$ (in Eq. (1)) by minimizing the cross-entropy loss of Eq. (1).

STA w/o FT. After pre-training ϕ by **Shuffle-Then-Assemble** strategy, the parameters of ϕ (in Eq. (2)) are fixed. When training the network by minimizing Eq. (1), only parameters of $MLP(\cdot)$ (in Eq. (1)) are updated.

STA w/o Res. The transformation network in Fig 3 is not a residual network. And the other settings are the same with STA.

We also compared with state-of-the-art visual relationship prediction methods such as **VTransE** [60], **Lu’s-V** [32], **Lu’s-VLK** [32], and **Peyre’s-A** [38]. Note that except for Lu’s-VLK which is a multimodal model, all the methods compared here are visual models.

4.3 Results and Analysis

Table 1, 2 show the performance of compared methods on two datasets of different experiment settings. As we can see, the proposed STA has the best performances compared with the other baselines and state-of-the-art on both datasets. For example, compared to the Base+OA, the proposed STA can boost the Recall@100 of supervised, weakly supervised, and zero-shot relationship prediction by absolute 4.75%, 4.42%, 4.04%, respectively on VRD, and 4.41%, 4.2%, 5.81%, respectively on VG.

Table 1: The performances (Recall@K%) of compared methods on two datasets under Supervised setting and Supervised (Det) setting.

Dataset	VRD		VG		VRD(Det)		VG(Det)	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
Base	39.25	39.25	52.48	52.61	37.83	37.83	50.12	50.31
Base+OA	43.29	43.29	58.35	58.53	40.78	40.78	57.03	57.31
STA w/o FT	44.30	44.30	58.14	58.32	41.12	41.12	56.88	57.02
STA w/o Res	46.83	46.83	62.08	62.32	44.85	44.85	61.12	61.30
STA	48.03	48.03	62.71	62.94	45.65	45.65	61.27	61.51
Lu’s-V [32]	7.11	7.11	–	–	–	–	–	–
Lu’s-VLK [32]	47.87	47.87	–	–	–	–	–	–
VTransE [60]	44.76	44.76	62.63	62.87	–	–	–	–
Peyre’s-A [38]	46.30	46.30	–	–	–	–	–	–

Table 2: The performances (Recall@K%) of compared methods on two datasets under Weakly Supervised setting (WS) and Zero-Shot setting(ZS).

Dataset	VRD(WS)		VG(WS)		VRD(ZS)		VG(ZS)	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
Base	29.36	29.36	45.78	46.01	14.10	14.10	11.04	11.04
Base+OA	31.47	31.47	47.46	47.72	16.53	16.53	13.09	13.09
STA w/o FT	32.84	32.84	47.23	47.39	18.24	18.24	13.72	13.72
STA w/o Res	35.10	35.10	50.89	51.13	19.01	19.01	18.03	18.03
STA	35.89	35.89	51.73	51.92	20.57	20.57	18.90	18.90
Peyre’s A [38]	34.03	34.03	–	–	16.10	16.10	–	–

Comparing the results of Base+OA with Base, we can see that by adding OA conv-layers, the performance is improved. This observation is basically as expected since the number of parameters have been increased and thus the representation ability of the whole network is improved. By comparing the performance of STA w/o FT with Base+OA, we can find that, even OA conv-layers are not fine-tuned, the features which are pre-trained by **Shuffle-Then-Assemble** still have comparable performance with the Base+OA. And when the pre-trained OA conv-layers are further fine-tuned (STA w/o Res, STA), the performances will have a considerable boost. Such observations show that the success of the proposed method is not only due to the added small network (OA conv-layers), but also thanks to the proposed **Shuffle-Then-Assemble** pre-training strategy.

Fig. 6 shows the R@100 of relationship prediction of the four relation types which are comparative, preposition, verb and spatial. From this, we can see that the proposed STA has the best performance in each relationship type on both datasets.

Analysis of feature maps. Fig. 5 shows six qualitative examples of feature maps generated by three different strategies. By comparing the STA’s feature maps with Base and Base+OA, we can find that STA’s feature maps focus more on the overlap regions between subjects and objects. For example, in the

Table 3: Computed overlap ratios (%) of two kinds of feature maps

Dataset	OA	Base CNN	Dataset	OA	Base CNN
VRD	50.27	42.45	VG	48.50	41.32

second row, STA’s feature maps put more attention on people’s feet, which would provide cues for predicting the right relationship “stand on”.

The ratio: $(\sum_{i \in R_{over}} f(i)) / (\sum_{i \in R_{joint}} f(i))$, in Table 3, is used to measure how our model can focus on the overlapped region. In this formula, $f(\cdot)$ is the normalized joint feature map of subject and object region, R_{over} and R_{joint} mean the overlapped region and the joint region of that feature map respectively. We compare the ratios computed by OA feature and Base CNN feature on both VRD and VG datasets. From the results we can see that the proposed **Shuffle-Then-Assemble** pre-training strategy can help the relationship model captures more attention on the shared regions between subject and object.

Analysis of Zero-Shot Setting. From table 2, we can see that the proposed STA has the best performance on both datasets compared with other baselines and one state-of-the-art. This result can further validate the effectiveness of the proposed **Shuffle-Then-Assemble** pre-training strategy. From the qualitative examples in Fig. 7, we can demonstrate that the reason why STA achieves better performance is due to the learned OA feature maps.

Analysis of object-biased relationships. Fig. 8 shows the accuracy of each relationship, listed in an ascending, left-right order according to their biases to specific subject-object configuration by $N_R(r)/N_C(r)$, where $N_C(r)$ is the number of configurations and $N_R(r)$ is the number of training samples of the r -th relationship. Notice that smaller bias indicates more flexible configurations (*e.g.*, “touch”) and vice versa (*e.g.*, “wear”). We can find that for relationships which are less biased to specific configurations (left and middle parts), our STA is better as it focuses on object-agnostic features.

Failure mode. Our model will fail when one relationship depends heavily on specific object combinations. For example, for some relationship listed in the right part of Fig. 8 (like the relationship “read”, the subjects and objects are usually “person” and “book”), our model will not defeat the baseline. Under this condition, the object categories will be useful for predicting relationship. Note that such failure can be easily recovered by rules mined from dataset statistics.

5 Conclusions

We proposed a novel **Shuffle-Then-Assemble** visual relationship feature learning strategy for improving visual relationship models. The key idea is to discard the original one-to-one paired object alignments, and then try to recover them in an unsupervised pair discovery fashion by using a cycle-consistent adversarial domain transfer method. In this way, the object class information in object pairs is excluded and hence the resultant feature map is less likely biased to

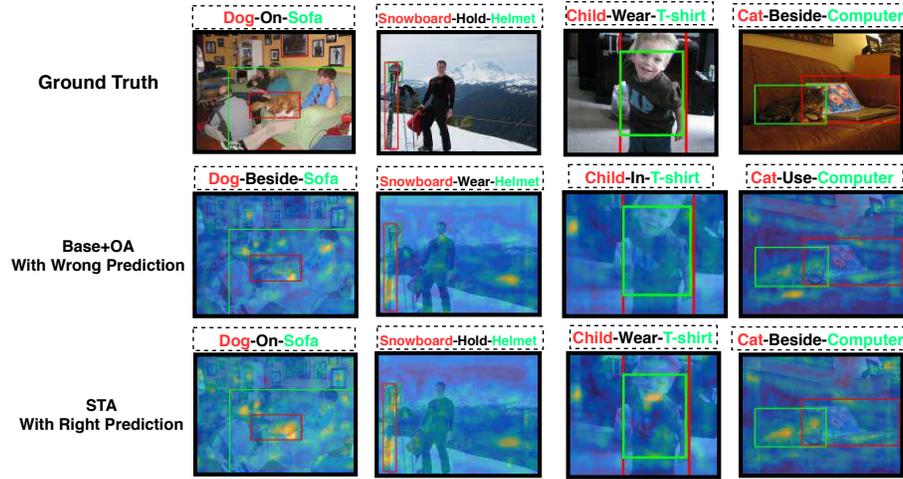


Fig. 7: Qualitative feature maps of four zero-shot relationships on VRD dataset. For each one, two feature maps of Base+OA with wrong prediction and STA with correct prediction are visualized by averaging over the 512 channels. We can see that by using the proposed **Shuffle-Then-Assemble** (STA), the ROI features are less likely biased to the objects and more focused on the regions of interaction of the two objects.

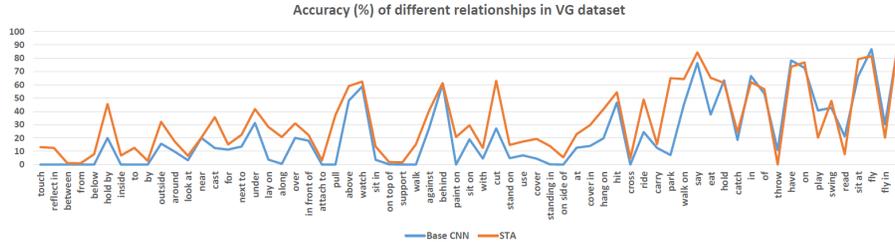


Fig. 8: The accuracy (%) of each relationship in VG dataset. In the horizontal axis, the relationships are listed in an ascending order (from left to right) of their biases to specific object combinations. The vertical axis is the accuracy (%) of each relationship. We can see that for relationships which are less biased to specific combination (left parts), our STA method usually have better performance.

specific object compositions. On two visual relationship benchmarks, we found a consistent improvement from a naive relationship prediction model using the pre-trained OA feature maps.

Acknowledgments. This research is partially support by NTU-CoE Grant, Alibaba-NTU JRI, and Data Science & Artificial Intelligence Research Centre@NTU (DSAIR).

References

1. Atzmon, Y., Berant, J., Kezami, V., Globerson, A., Chechik, G.: Learning to generalize to new compositions in image understanding. In: EMNLP (2016)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. TPAMI (2013)
3. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: ICCV (2015)
4. Chen, L., Zhang, H., Xiao, J., Liu, W., Chang, S.F.: Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In: CVPR (2018)
5. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR (2017)
6. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: CVPR (2017)
7. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
9. Desai, C., Ramanan, D.: Detecting actions, poses, and objects with relational phraselets. In: ECCV (2012)
10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. pp. 1422–1430 (2015)
11. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning (2017)
12. Girshick, R.: Fast r-cnn. In: ICCV (2015)
13. Gu, J., Cai, J., Wang, G., Chen, T.: Stack-captioning: Coarse-to-fine learning for image captioning. In: AAAI (2018)
14. Gu, J., Wang, G., Cai, J., Chen, T.: An empirical study of language cnn for image captioning. In: ICCV (2017)
15. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recognition (2017)
16. Gupta, A., Davis, L.S.: Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In: ECCV (2008)
17. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. TPAMI (2009)
18. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. CVPR (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
21. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR (2017)
22. Jabri, A., Joulain, A., van der Maaten, L.: Revisiting visual question answering baselines. In: ECCV (2016)
23. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: CVPR (2015)

24. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML (2017)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
27. Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. arXiv preprint arXiv:1803.07464 (2018)
28. Li, Y., Ouyang, W., Wang, X., et al.: Vip-cnn: Visual phrase guided convolutional neural network. In: CVPR (2017)
29. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
31. Liu, D., Zha, Z.J., Zhang, H., Zhang, Y., Wu, F.: Context-aware visual policy network for sequence-level image captioning. In: ACMMM (2018)
32. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV (2016)
33. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NIPS. pp. 405–415 (2017)
34. Ma, L., Sun, Q., Georgoulis, S., Gool, L.V., Schiele, B., Fritz, M.: Disentangled person image generation. In: CVPR (2018)
35. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: ICCV (2017)
36. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV (2016)
37. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
38. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: ICCV (2017)
39. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)
40. Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rosenberg, C., Fei-Fei, L.: Learning semantic relationships for better action retrieval in images. In: CVPR (2015)
41. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR (2016)
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
43. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR (2011)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
45. Sun, Q., Ma, L., Oh, S.J., Gool, L.V., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: CVPR (2018)
46. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
47. Tao, Q., Yang, H., Cai, J.: Zero-annotation object detection with web knowledge transfer. arXiv preprint arXiv:1711.05954 (2017)
48. Tsai, Y.H., Hung, W.C., Schuler, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)

49. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. TPAMI (2017)
50. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: CVPR (2018)
51. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
52. Xu, D., Zhu, Y., Choy, C., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
53. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
54. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
55. Yatskar, M., Zettlemoyer, L., Farhadi, A.: Situation recognition: Visual semantic role labeling for image understanding. In: CVPR (2016)
56. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
57. Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., Loy, C.C.: Zoom-net: Mining deep feature interactions for visual relationship recognition (2018)
58. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS (2014)
59. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR (2018)
60. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: CVPR (2017)
61. Zhang, H., Kyaw, Z., Yu, J., Chang, S.F.: Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In: ICCV (2017)
62. Zhang, H., Niu, Y., Chang, S.F.: Grounding referring expressions in images by variational context. In: CVPR (2018)
63. Zhang, J., Elhoseiny, M., Cohen, S., Chang, W., Elgammal, A.: Relationship proposal networks. In: CVPR (2017)
64. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: CVPR (2017)
65. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
66. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
67. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: ICCV (Oct 2017)