DEEPLY SUPERVISED MULTIMODAL ATTENTIONAL TRANSLATION EMBEDDINGS FOR VISUAL RELATIONSHIP DETECTION

Nikolaos Gkanatsios[1,2], Vassilis Pitsikalis[1], Petros Koutras[2], Athanasia Zlatintsi[2], Petros Maragos[2]

[1]DeepLab, 11635, Athens, Greece
[2]School of E.C.E., National Technical University of Athens, 15773, Athens, Greece
Email: nikos.gkanatsios93@gmail.com, vpitsik@deeplab.ai, {pkoutras, nzlat, maragos}@cs.ntua.gr

ABSTRACT

Detecting visual relationships, i.e. <Subject, Predicate, Object> triplets, has been a challenging Scene Understanding task approached in the past via linguistic priors or spatial information in a single feature branch. We introduce a new deeply supervised two-branch architecture, the Multimodal Attentional Translation Embeddings, where the visual features of each branch are driven by a multimodal attentional mechanism that exploits spatio-linguistic similarities in a low-dimensional space. We present a variety of experiments comparing against all related approaches in the literature, as well as by re-implementing and fine-tuning several of them. Results on the commonly employed VRD dataset [1] show that the proposed method clearly outperforms all others, while we also justify our claims both quantitatively and qualitatively.
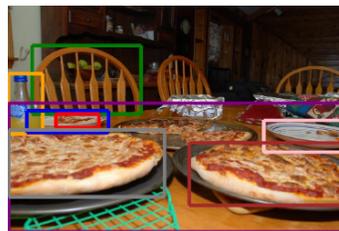
Index Terms— Visual Relationship Detection, Attention, Multiple Modalities, Translation Embeddings, Deep Supervision.

1. INTRODUCTION

Is this "pizza slice *on* the plate" or "*next to* the plate?" The classification of such semantic visual relationships, namely Visual Relationship Detection, aims on bridging the gap between visual and semantic perception by detecting $< S, P, O >$ triplets, with the predicate P being the relationship of subject S and object O, e.g. "pizza - on - plate" (Fig. 1). The large intra-class variance of the predicates' appearance and the vast number of unbalanced classes led prior works to learn S, P and O separately and employ linguistic [1, 2] and spatial information [3, 4] to deal with the predicates' visual variance. Recently, [5] proposed Visual Translation Embeddings (VTransE), a representation of relationships as a vector translation, e.g. cat + eat ≈ fish. Albeit competent, VTransE uses only subject and object features and no predicate or contextual information, proved effective lately [2, 6].

Inspired by VTransE [5], the complementary nature of language and vision [1] and the recent advances in attention mechanisms [7] and Deep Supervision [8], we introduce Multimodal Attentional Translation Embeddings (MATransE): 1) We directly use predicate features and subject-object features, in two separate branches, 2) we employ a spatio-linguistic attention mechanism to drive the visual features, 3) we perform an alignment of the two branches in a common score space using Deep Supervision. Our system is

Fig. 1: Visual Relationship Detection refers to localizing pairs of objects (S, P) and classifying their interactions P to form $< S, P, O >$ triplets. The exponential number of possible such relationships and the visual variability of P – e.g. "behind" on "bottle - behind - pizza" and "chair - behind - table" – render the task quite challenging.

evaluated on the widely used VRD [1] dataset. We also shed light to inconsistencies in the metric's definition that made prior works incomparable and re-implement and fine-tune several baselines so as to be comparable. Our system outperforms all other methods, showing the importance of the proposed approach.

2. RELATED WORK

Visual Relationships Detection [9] was only recently formulated [1]. Most works detect objects [10] and then predict the predicate class of each object pair [1, 2, 3, 4]. We cluster the related literature based on the feature type and the fusion level of different modalities.

Visual Appearance Features are extracted from the predicate box, i.e. the minimum rectangle that encompasses the subject box and the object box [1, 11, 2, 12, 13, 3], the separate subject-object boxes [5, 14, 15, 16], or both [17, 18, 4, 19, 20]. All the above train a single branch with visual features, while we jointly train two separate branches with different features, a predicate feature branch (P-branch) and an object-subject branch (OS-branch), and employ Deep Supervision to align their scores into a common space.

Linguistic and Semantic Features are employed in a feature-level integration with word embeddings [1, 12, 4, 11, 3], encoding of statistics [17, 12, 4, 13], late-fusion with subject-object classemes (score vectors) [21, 5, 13, 18] and loss-level fusion as regularization terms [1, 3] or adaptive-margins [4, 18, 11]. Closest to us, [2] uses subject-object embeddings to train context-aware classifiers and [19] trains multimodal embeddings by projecting visual and linguistic features into a common space. Different from these approaches, we use language as a component of our multimodal attentional scheme to drive the visual features of the two branches.

Lastly, *Spatial Features*, either hand-crafted [3, 12, 13, 16] or
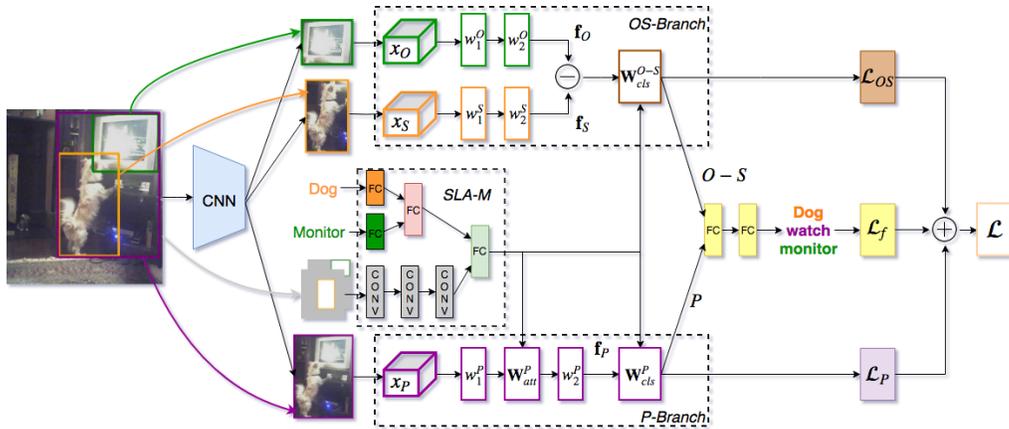
Fig. 2: Overview of the proposed approach. Given a detected subject-object pair, MATransE employs spatial and linguistic information (binary masks and word embeddings) in an attention module (SLA-M) that guides the classification of the convolutional features of the subject (orange), object (green) and predicate box (purple). Subject-object and predicate features are handled into two separate branches, OS-Branch and P-Branch respectively, and their scores are fused. The single branches and their fusion all contribute to the total loss (terms $\mathcal{L}_{\mathbf{OS}}, \mathcal{L}_{\mathbf{P}}, \mathcal{L}_{\mathbf{f}}$) and are jointly optimized with Deep Supervision, forcing an alignment in the score space such that $P \approx O - S$.

convolutional [21, 4, 18, 11], are often integrated in a feature- [21, 4, 18, 11], score- [20] or loss-level fusion [3]. Recently, [22] used masks as an early-stage attention mechanism to detect objects related to a given one. We integrate mask features into our spatio-linguistic attention scheme to control visual features' classification.

Visual Translation Embeddings [5] is the most related work, mapping S, P and O in a vector space where valid relationships satisfy $S + P \approx O$. Nonetheless, no predicate features are used, leaving the alignment of the two terms of the equation to take place inside the loss function. Instead, our architecture is two-branch and its components are jointly optimized using Deep Supervision. Predicate and subject-object features are explicitly used, guided by a spatio-linguistic attention mechanism.

3. METHODOLOGY

3.1. Background

Visual Relationship Detection involves simultaneous detection of all $< S, P, O >$ triplets in a given image by localizing S and O and classifying their interaction P. In addition to the high visual variability of S, P, O, another challenge is the large number of possible triplets, e.g. 100 object and 70 predicate classes yield 700000 relationships. We adopt an approach scalable to thousands of relationships by learning to predict each component separately [1], factorizing the Visual Relationship Detection process as:

$$Pr(< S, P, O > |I) = Pr(S, O|I)Pr(P|S, O, I), \quad (1)$$

separating object detection (1st term) from predicate classification (2nd term). The first can be addressed by a state-of-the-art object detector like [10, 23], so the rest of this paper focuses on the more challenging second one. For this term, [5] exploits features from the detected subject and object and models relationships by projecting S, P and O into an embedding space where:

$$S + P \approx O \quad (2)$$

Following this formulation, we introduce Multimodal Attentional Translation Embeddings, MATransE, that guides the features' projection with attention and Deep Supervision to satisfy Eq. (2).

3.2. Multimodal Attentional Translation Embeddings

MATransE learns a projection of $< S, P, O >$ into a score space where $S + P \approx O$, by jointly representing S, P and O as $\mathbf{W}_S x_S$, $\mathbf{W}_P x_P$ and $\mathbf{W}_O x_O$. Vectors x_S, x_P, x_O are visual appearance features [24], extracted both from the predicate image (union of subject and object bounding boxes) and the subject and object images of a pair of detected objects. To learn the projection matrices $\mathbf{W}_S, \mathbf{W}_P, \mathbf{W}_O$, MATransE employs a Spatio-Linguistic Attention module (SLA-M) that uses binary masks' convolutional features m [21] and encodes subject and object classes with pre-trained word embeddings s, o [25, 1]. Therefore, Eq. (2) becomes:

$$\mathbf{W}_P(s, o, m)x_P = \mathbf{W}_O(s, o, m)x_O - \mathbf{W}_S(s, o, m)x_S \quad (3)$$

This allows us to design a two-branch architecture: a branch driving the predicate features into scores $\mathbf{W}_P(s, o, m)x_P$ (P-branch) and another classifying the difference of weighted object-subject features $\mathbf{W}_O(s, o, m)x_O - \mathbf{W}_S(s, o, m)x_S$ (OS-branch). To satisfy Eq. (3), we enforce a score-level alignment by jointly minimizing the loss of each one of the P- and OS-branch with respect to ground-truth using Deep Supervision [8]. See Fig. 2 for an overview.

Spatio-Linguistic Attention Module (SLA-M) combines a 3-layer CNN [21] to encode binary mask features and a MLP to encode subject and object embeddings. The spatial and linguistic information are fused with a fully-connected (FC) layer that performs a projection of the spatio-linguistic vector into a low-dimensional (64-d) space where spatially and semantically similar configurations come closer. For instance, "man on horse" might be spatio-linguistically closer to "man ride elephant" than "man next to horse" and "cat on horse" that share either the same linguistic, i.e. "man", "horse", or spatial, i.e. "on", information.

Predicate Branch: Predicate images often contain irrelevant information that "distracts" the network, thus, we use pre-pooled predicate features that have spatial dimensions as well and apply attentional pooling [2] to concentrate the network's "focus" on the discriminative visual cues that may only appear in a small fraction of the image. Intuitively, the attention function should be different depending on the objects' classes and their spatial configuration. Thus,

our attentional pooling weights $\mathbf{W}_{att}^{(P)}(s, o, m)$ are directly dependent on SLA-M's output.

The pooled feature vector is further encoded via a FC layer and is classified into predicate scores using an attentional classifier. Once again, it is easier to classify the predicate knowing that the relationship concerns certain objects at specific relative position and sizes. We thus obtain the P-branch scores as:

$$P = \mathbf{f}_P \mathbf{W}_{cls}^{(P)}(s, o, m) + \mathbf{b}_{cls}^{(P)} \qquad (4)$$

with $\mathbf{f}_P$ the encoded predicate feature vector, $\mathbf{W}_{cls}^{(P)}(s, o, m)$ the spatio-linguistic classifier weights and $\mathbf{b}_{cls}^{(P)}$ a bias vector (Fig. 2).

Object-Subject Branch: Subject and object regions are far less noisy and we therefore consider pooled feature vectors to be a meaningful representation. Contrary to [5], we do not subtract the subject and object features directly but we first encode them as $\mathbf{f}_S$, $\mathbf{f}_O$ using two FC layers, as the subject and object regions often occlude each other, so that pure subtraction would possibly weaken meaningful features. The difference of the encoded features is classified using an attentional classifier, similar to the P-branch:

$$O - S = (\mathbf{f}_O - \mathbf{f}_S) \mathbf{W}_{cls}^{(O-S)}(s, o, m) + b_{cls}^{(O-S)} \qquad (5)$$

with $\mathbf{W}_{cls}^{(O-S)}(s, o, m)$ denoting the spatio-linguistic attentional classifier weights and $b_{cls}^{(O-S)}$ a bias vector.

Fusion with Deep Supervision: We combine P- and OS-branches' scores into a single vector and train a meta-classifier to obtain the predicate classes, minimizing the fusion loss $\mathcal{L}_f$. The scores of each branch are deeply supervised, i.e. we demand that they give a meaningful prediction as well, minimizing $\mathcal{L}_P$ and $\mathcal{L}_{OS}$ respectively. Thus, with $\mathbf{W} = (\mathbf{W}_P, \mathbf{W}_O, \mathbf{W}_S)$, the total loss:

$$\mathcal{L}(\mathbf{W}) = \lambda_f \mathcal{L}_f(\mathbf{W}) + \lambda_P \mathcal{L}_P(\mathbf{W}_P) + \lambda_{OS} \mathcal{L}_{OS}(\mathbf{W}_O, \mathbf{W}_S) \quad (6)$$

where λ hyperparameters balance each term's importance. The optimization problem is solved using Adam optimizer [26].

4. EVALUATION AND RESULTS

4.1. Dataset and Metrics

We evaluate our approach on the widely benchmarked VRD [1]; this contains 5000 images with 100 object categories, 70 predicates, 37993 annotated relationships and 6672 relationship types. We report results for Predicate Detection, a task that directly quantifies Visual Relationship Detection. In this setup, the objects' boxes and classes are given and the task is to predict the predicate.

As explained in [1], a suitable metric is $\mathrm{Recall}@x$, that counts the fraction of times the correct relationship is included in the top x confident predictions, where $x = 50, 100$. However, there is one hyperparameter k in the metric computation that prior works often do not specify, leading to unfair comparisons. To address this, we re-formulate the metric as $\mathrm{Recall}_k@x$ ($R_k@x$). Let N be the number of examined subject-object pairs in an image; then, keeping the top-k predictions per pair, $R_k@x$ examines the x most confident predictions out of Nk total.

Most works have seen Predicate Detection as a multiclass problem and they use $k = 1$ to reward the correct top-1 prediction for each pair [1, 16, 3, 5, 2]. Motivated by the fact that there are pairs annotated with more than one predicate classes, other works [4, 11] have tackled this as a multilabel problem and they use $k = 70$ to allow for predicate co-occurrences [4, 11]. Intuitively, the two choices do not always converge to the same optimum, as the one favors only

Method	R_1@50	R_{70}@50	R_{70}@100
VTransE [5]	44.7	-	-
PPR-FCN [27]	47.7	-	-
VRD [1]	47.8	-	-
STA [15]	48.03	-	-
S-R [3]	51.5	-	-
Weak Supervision [16]	52.6	-	-
Sem. Modeling [17]	53.1	-	-
CAI [2]	53.59	-	-
Guided Proposals [14]	-	71.3	81.8
DR-Net [21]	-	80.7	81.9
ROR [12]	-	-	82.1
Bi-RNN [28]	-	84.9	92.6
DSR [4]	-	86.0	93.1
CDDN [11]	-	87.57	93.76
Zoom-Net [6]	50.6	84.2	90.5
LK (VRD+VG) [13]	54.82	83.97	90.63
Zoom-Net + CAI [6]	<u>55.9</u>	<u>89.0</u>	<u>94.5</u>
VTransE (ours)	46.2	78.2	87.8
VRD (ours)	47.8	81.7	91.1
CAI (ours)	51.0	85.69	93.8
ROR (ours)	53.0	82.8	92.1
DR-Net (ours)	53.79	79.47	91.31
Ours MATransE	**56.14**	**89.79**	**96.26**

Table 1: Comparisons on VRD Predicate Detection with all related reported and re-implemented ("(ours)") results. Bold fonts indicate best performance and underlined ones the 2nd best. Note that $R_1@50 = R_1@100$ [1] on VRD, thus we report only $R_1@50$.

the most probable class, while the other does not force any of the true labels to be the top prediction, leaving prior works incomparable due to their different viewpoint of the task itself.

4.2. Experimental results

We first compare our method against reported as well as re-implemented results of past approaches (Table 1), then we further discuss our system's components (Table 2). Our PyTorch code is publicly available[1].
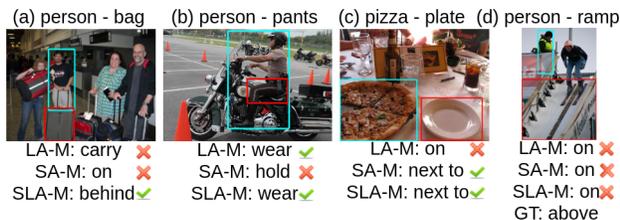
Quantitative Results: To address the ambiguity in the metric that prior works have used (Sec. 4.1), we sort the methods based on k and report results for all such choices. To the best of our knowledge, we are the first to review all past works that report Predicate Detection results on VRD, see Table 1, where it is clear that our method outperforms prior works in all three metric settings, without the need of external corpora [13] or fine-tuning of the feature extractor [2]. Specifically, comparisons using $R_1@50$ show that we achieve 0.43%, 2.41% and 4.76% respective relative improvement over the main competitors [6, 13, 2]. Using $R_{70}@100$, the main competitors are different [6, 11, 4] but still MATransE achieves 32%, 40.1% and 45.8% relative error decrease.

To extend the comparisons to the literature even further, we re-implement and fine-tune several approaches previously not comparable. In most cases, fine-tuning boosted the results over the originally reported. Moreover, we validate that the different metric settings do not always converge to the same solution. An interesting observa-
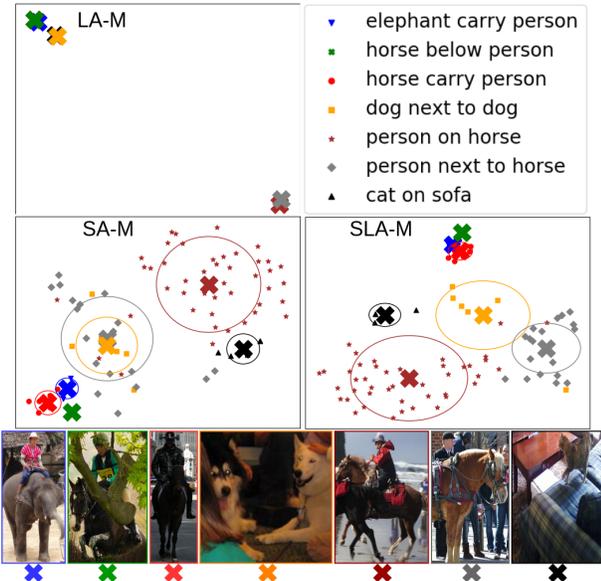
Fig. 3: Visualization [29] of the projection space created by each attentional module; 7 classes. Centroids and std ellipsoids are also shown. LA-M quantizes relationships based exclusively on subject and object embeddings. SA-M generates weights that are highly-dependent on the objects relative positions. SLA-M exploits both spatial and semantic features to create a space where different relationships are better separated, while preserving the desired similarities. Best viewed in color.

Fig. 4: Qualitative results using different attention modules are reported below the corresponding image. Ticks and crosses mark correct and incorrect results respectively. Subject and object boxes are in cyan and red respectively. (d) is a failure case (GT: ground-truth).

Method	$R_1@50$	$R_{70}@50$	$R_{70}@100$
(P + OS) + DS	46.43	82.36	92.03
LA-M + (P + OS) + DS	54.56	88.31	95.59
SA-M + (P + OS) + DS	49.83	85.46	93.39
SLA-M + P	55.82	89.6	95.97
SLA-M + OS	55.85	89.72	96.22
SLA-M + (P + OS)	55.14	88.29	95.06
SLA-M + (P + OS) + DS	**56.14**	**89.79**	**96.26**

Table 2: Experimental comparisons by removing one component at a time. We get best results with the full system, that combines the Spatio-Linguistic Attention module (SLA-M) and P- and OS-branches, jointly trained with Deep Supervision (DS), outperforming the linguistic (LA-M) or spatial attention (SA-M) or single branches.

to 0.187 when Deep Supervision is used and 0.722 otherwise. In the latter case, redundant parameters obstruct convergence to an optimal solution, proving the importance of MATransE's formulation.

Qualitative Results: Figure 3 is a visualization [29] of the projection space created by each attentional module. Since LA-M exploits only subject and object embeddings, it quantizes the vector space and generates similar attentional weights for semantically similar object pairs, e.g. "horse - next to - person" and "elephant - carry - person". On the other hand, SA-M generates weights that are highly-dependent on the objects' positions and forms three "super-clusters": "carry - below", "next to" and "on". SLA-M combines the benefits of both modules and creates a space where the different relationships are separated, while preserving the desired neighborhoods. For instance, "dog - next to dog" is far from "person - next to - horse", a benefit of linguistic similarity, "person - next to - horse" is far from "person - on - horse", a benefit of spatial similarity and "horse - below - person" is very close to "elephant - carry - person", a benefit of both spatial and linguistic similarity. Note that in Fig. 3, SA-M has projected "dog - next to - dog" very close to "person - next to - horse", while this is resolved in SLA-M projections.

We notice a similar behavior on Fig. 4, where the system that employs LA-M tends to predict the most semantically common relationships, e.g. it is more common for a pizza to be on a plate than next to it. SA-M favors predictions that share spatial configurations, e.g. in Fig. 4(b) the relative position of subject and object bounding boxes is such that SA-M favors "hold". SLA-M is more robust than both SA-M and LA-M but may fail when both spatial and semantic information are "distractive" (Fig. 4(d)).

5. CONCLUSION

We address the challenging task of Visual Relationship Detection introducing MATransE, a novel deeply supervised network that uses spatio-linguistic attention to drive the features of two branches and aligns their scores into a common space. Our experiments prove the significance of MATransE's components, by achieving the best scores on VRD dataset under all different metric settings. A future direction would be to test our method on the newly introduced and larger Visual Genome [30]. Lastly, we show that our spatio-linguistic module creates a projection space where relationship triplets are represented as multimodal embeddings with desired properties. Their comparison with language models could be investigated to create knowledge representation priors using spatio-linguistic embeddings.

tion, for instance, is that although [21] largely outperforms [1] based on $R_1@50$, their $R_{70}@100$ performance is very similar. Even after improving the other methods' results, we still achieve better performance than all re-implemented ones.

Spatio-Linguistic Attention: We test our network's performance by totally removing SLA-M or by replacing it with LA-M (linguistic) or SA-M (spatial) attention module. All attentional modules clearly outperform the case without attentional mechanism, as they lead the network's "focus" on the most discriminative features (Table 2). Moreover, SLA-M clearly outperforms both LA-M, SA-M.

Two-branch architecture and Deep Supervision: Although competent, none of the single branches alone performs on par with the jointly trained 2-branch MATransE (Table 2). This is not true when Deep Supervision is not used and the two branches are merged, as removing $\mathcal{L}_P$ and $\mathcal{L}_{OS}$ from Eq. (6) cancels Eq. (3): we measure the mean value of $\|P - (O - S)\|_2$ on the test set and find it equal

6. REFERENCES

[1] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei, "Visual Relationship Detection with Language Priors," in *Proc. ECCV*, 2016.

[2] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian D. Reid, "Towards Context-Aware Interaction Recognition for Visual Relationship Detection," in *Proc. ICCV*, 2017.

[3] Yaohui Zhu, Shuqiang Jiang, and Xiangyang Li, "Visual Relationship Detection with Object Spatial Distribution," in *Proc. ICME*, 2017.

[4] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen, "Visual Relationship Detection With Deep Structural Ranking," in *Proc. AAAI*, 2018.

[5] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua, "Visual Translation Embedding Network for Visual Relation Detection," in *Proc. CVPR*, 2017.

[6] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy, "Zoom-Net: Mining Deep Feature Interactions for Visual Relationship Recognition," in *Proc. ECCV*, 2018.

[7] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip Torr, "Learn to Pay Attention," in *Proc. ICLR*, 2018.

[8] Saining Xie and Zhuowen Tu, "Holistically-Nested Edge Detection," in *Proc. ICCV*, 2015.

[9] Mohammad Amin Sadeghi and Ali Farhadi, "Recognition Using Visual Phrases," in *Proc. CVPR*, 2011.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," in *Proc. NIPS*, 2015.

[11] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang, "Context-Dependent Diffusion Network for Visual Relationship Detection," in *Proc. ACMMM*, 2018.

[12] Yusuke Goutsu, "Region-Object Relevance-Guided Visual Relationship Detection," in *Proc. BMVC*, 2018.

[13] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis, "Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation," in *Proc. ICCV*, 2017.

[14] Franois Plesse, Alexandru Ginsca, Bertrand Delezoide, and Franoise J. Prêteux, "Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation," in *Proc. ICME*, 2018.

[15] Xu Yang, Hanwang Zhang, and Jianfei Cai, "Shuffle-Then-Assemble: Learning Object-Agnostic Visual Relationship Features," in *Proc. ECCV*, 2018.

[16] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic, "Weakly-Supervised Learning of Visual Relations," in *Proc. ICCV*, 2017.

[17] Stephan Baier, Yunpu Ma, and Volker Tresp, "Improving Visual Relationship Detection Using Semantic Modeling of Scene Descriptions," in *Proc. ISWC*, 2017.

[18] Yaohui Zhu and Shuqiang Jiang, "Deep Structured Learning for Visual Relationship Detection," in *Proc. AAAI*, 2018.

[19] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed M. Elgammal, and Mohamed Elhoseiny, "Large-Scale Visual Relationship Understanding," *CoRR*, vol. abs/1804.10660, 2018.

[20] Ji Zhang, K. Shih, Andrew Tao, Bryan Catanzaro, and Ahmed Elgammal, "An Interpretable Model for Scene Graph Generation," *CoRR*, vol. abs/1811.09543, 2018.

[21] Bo Dai, Yuqi Zhang, and Dahua Lin, "Detecting Visual Relationships with Deep Relational Networks," in *Proc. CVPR*, 2017.

[22] Alexander Kolesnikov, Christoph H. Lampert, and Vittorio Ferrari, "Detecting Visual Relationships Using Box Attention," *CoRR*, vol. abs/1807.02136, 2018.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.03385, 2015.

[25] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, vol. abs/1301.3781, 2013.

[26] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[27] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang, "PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN," in *Proc. ICCV*, 2017.

[28] Wentong Liao, Shuai Lin, Bodo Rosenhahn, and Michael Ying Yang, "Natural Language Guided Visual Relationship Detection," *CoRR*, vol. abs/1711.06032, 2017.

[29] Laurens van der Maaten and Geoffrey E. Hinton, "Visualizing Data Using t-SNE," *JMLR*, vol. 9, 2008.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *IJCV*, vol. 123, 2016.