# From Saturation to Zero-Shot Visual Relationship Detection Using Local Context

Nikolaos Gkanatsios[1,2]
ngan@deeplab.ai
Vassilis Pitsikalis[1]
vpitsik@deeplab.ai
Petros Maragos[2]
maragos@cs.ntua.gr

[1] deeplab.ai
Athens, Greece
[2] National Technical University of Athens
Athens, Greece

**Abstract**

Visual relationship detection has been motivated by the "insufficiency of objects to describe rich visual knowledge". However, we find that training and testing on current popular datasets may not support such statements; most approaches can be outperformed by a naive image-agnostic baseline that fuses language and spatial features. We visualize the errors of numerous existing detectors, to discover that most of them are caused by the coexistence and penalization of antagonizing predicates that could describe the same interaction. Such annotations hurt the dataset's causality and models tend to overfit the dataset biases, resulting in a saturation of accuracy to artificially low levels.

We construct a simple architecture and explore the effect of using language on generalization. Then, we introduce adaptive local-context-aware classifiers, that are built on-the-fly based on the objects' categories. To improve context awareness, we mine and learn predicate synonyms, i.e. different predicates that could equivalently hold, and apply a distillation-like loss that forces synonyms to have similar classifiers and scores. The last also serves as a regularizer that mitigates the dominance of the most frequent classes, enabling zero-shot generalization. We evaluate predicate accuracy on existing and novel test scenarios to display state-of-the-art results over prior biased baselines.

## 1 Introduction

Integrating visual relationships, i.e. <*S*ubject, *P*redicate, *O*bject> triplets, on a directed graph structure has powered applications such as image retrieval [15] and generation [11, 16, 30, 43], captioning [22, 50] and visual question answering [8, 53, 59]. Scene graphs' significance springs from the observation that images are more than disconnected objects together [21, 28] (Fig. 1a). Results on the most commonly used datasets [20, 28], however, show that a baseline relying only on object categories and locations is able to outperform most state-of-the-art methods (Fig. 1b). On the other hand, heavy use of such priors [3, 6, 57, 58, 63] leads to saturated performance and misclassification of the same examples (Fig. 2).

A major cause for this behavior is the severe lack of causality in the datasets' annotations, due to the ambiguous and overlapping predicate meanings, which we call *synonyms* (Fig. 1a). Having multiple synonyms antagonizing each other with a standard Cross-Entropy

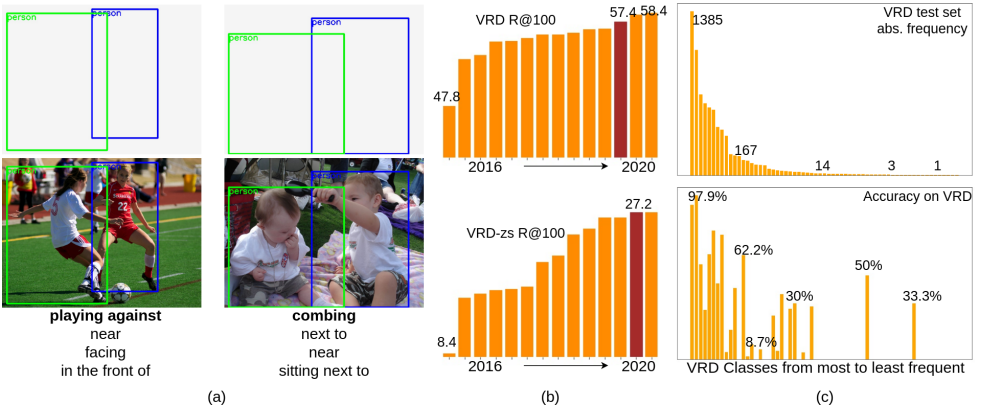


Figure 1: (a) A rich scene representation demands more than objects' categories and locations. (b) However, an image-agnostic baseline based on such features, highlighted with darker color on the recall (R@100) bar plots, still prevails over most methods, even in zero-shot cases (VRD-zs). (c) If we examine per-class accuracy, we find that such models learn to mimic the dataset's biases: the co-existence of multiple antagonizing predicates with the same meaning, e.g. "next to", "near" and "siting next to" on subfigure (a), hurt causality in favor of the most frequent predicates per local context (subject-object categories).

loss (CE) and a single ground-truth confuses networks towards the most generic and frequent predicates (Fig. 1c), hampering potential higher-level applications [42]. To strengthen this, we implement several models, to find that their predictions are highly correlated and often wrongly penalized as incorrect (Sec. 2 and Fig. 2). For instance, in Fig. 2c, four models predict "next to", which is the most frequent synonym for the ground-truth "by". The fact that the annotators interchangeably use these predicates for the same data perturbs by construction their semantic information. Even worse, as we further discuss in Sec. 2, in order to answer "by", such models *have to forget* "next to".

The above observations question the validity of current evaluation protocols. Indeed, if we merge synonyms and re-evaluate, we unfold a large gain for all tested models on three datasets (see Sec. 2), accompanied by an increased correlation of their outputs (Fig. 2e). Such results suggest that the true margin for improvement on current setups remains unclear.

The goal of this work is to explain the effect of language and synonyms on performance, challenge training and testing on existing datasets and reclaim the importance of contextual visual cues towards zero-shot generalization. We start with an analysis on common datasets and test multiple baselines on different setups, to show that more than half of the gap between current state-of-the-art performance and the upper bound of 100% is due to synonym predicates, concealing other causes of errors. Next, we build on visual and spatial features and study the effect of integrating linguistic attention. We boost our network's performance by introducing a recurrent local-context-aware classifier, that learns to build predicate templates on-the-fly, conditioned on the object categories. To minimize competition between predicate synonyms, we propose a loss term that pulls their classifiers closer, improving the learning of rarer classes by considering their similarity with more common classes. We also examine a regularization of the output space using a local-context-wise smoothed entropy-loss that distills semantic knowledge from a synonym-aware teacher. Lastly, we further contribute with

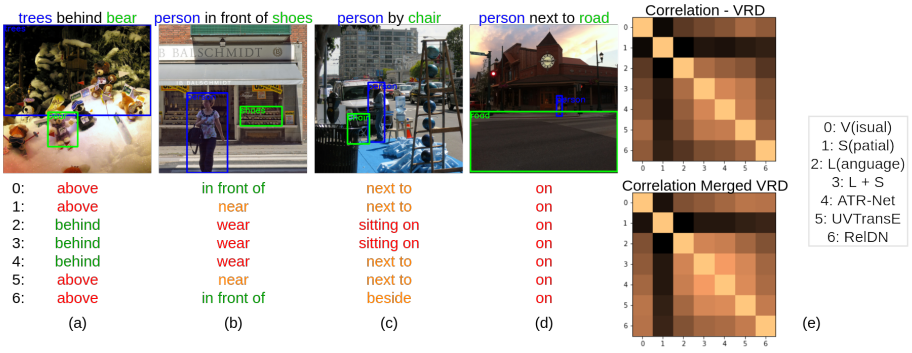


Figure 2: (a-d) Predictions of the seven baselines of Sec. 2 on different images. Green fonts denote correct classification, red erroneous and orange correct synonyms that are penalized as wrong. (e) Correlation between models' outputs for standard (above) and merged (below) VRD annotations. Save the spatial model, the outputs of all other models, especially those who use language, are highly correlated, inclining to misclassify the same samples.

benchmarking novel evaluation scenarios that are able to demonstrate the effectiveness of zero-shot classifiers. Comparisons on these setups showcase clear margins of the proposed method against other approaches, achieving a new state-of-the-art on zero-shot VRD [28] and UnRel [32] while remaining on par with the state-of-the-art on the full set.

## 2 Saturation on Visual Relationship Detection Datasets

We present an analysis on VRD [28] and Visual Genome (VG) [20] to showcase and explicate the saturation of current methodologies' results. Since there are more than 10 VG variants used in recent literature [9], we choose the most commonly used VG200 [48] and VrR-VG [27] that is constructed to focus on visually-relevant relations. To the best of our knowledge, we are the first to report zero-shot detection results on VG regardless of split.

We implement four baselines that use only visual (*V*), linguistic (*L*), spatial (*S*) and a combination of linguistic and spatial (*LS*) features. Next, we re-implement three state-of-the-art models, *ATR-Net* [9], *UVTransE* [14] and *RelDN* [63], that exploit all kinds of the aforementioned features in diverse ways. Despite the vastly different architectures, all these models display saturation on a seemingly low threshold of 58.5% on VRD, 70.2% on VG200 and 54.5% on VrR-VG, while *LS* performs on par with the state-of-the-art (Table 1).

We exploit the fact that VRD and VG often have multiple predicates annotated for a single sample to mine synonyms: if the same sample (pair of objects) is annotated with more than one predicates, then these predicates should be synonyms given these subject and object labels. A common example is "person has jacket" and "person wears jacket", that are often annotated together. We do not proceed however into saying that "has" is generally equivalent to "wears", respecting polysemy of predicate words in different context. This allows us to *merge* labels that represent a group of synonym predicates under specific context. Note that these labels are used exclusively for evaluation (denoted with "merged" in Tables 1 and 2): all models are trained on standard VRD/VG annotations.

*Considering predicate synonyms erases a great percentage of errors* on all three datasets,

| Model | VRD | | VG200 | | VrR-VG | | VRD zs | | VG200 zs | | VrR-VG zs | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | top-1 | merged | top-1 | merged | top-1 | merged | top-1 | merged | top-1 | merged | top-1 | merged |
| *V* | 52.59 | 77.00 | 64.55 | 89.01 | 45.22 | 69.49 | 24.57 | 36.47 | *23.68* | *31.81* | 25.79 | 36.91 |
| *L* | 54.28 | 79.55 | 68.64 | 91.58 | 52.54 | 74.43 | 17.12 | 28.77 | 15.44 | 24.14 | 25.31 | 36.42 |
| *S* | 47.58 | 70.49 | 52.6 | 83.21 | 25.68 | 57.67 | 25.51 | 36.82 | 19.48 | 27.14 | 17.48 | 28.46 |
| *LS* | *57.56* | *82.06* | *69.67* | *92.20* | ***54.52*** | ***75.65*** | 26.37 | 39.55 | 19.57 | 28.52 | 28.28 | 39.5 |
| *ATR-Net* [9] | ***58.48*** | ***82.76*** | ***70.18*** | ***92.42*** | 54.43 | *75.44* | *27.10* | *40.50* | 22.31 | 31.35 | *28.64* | *39.99* |
| *UVTransE* [14] | 57.25 | 81.50 | 69.05 | 91.72 | *54.50* | 75.43 | ***27.48*** | ***40.58*** | ***24.69*** | ***33.37*** | ***31.21*** | ***42.32*** |
| *RelDN* [63] | 54.47 | 78.90 | 67.11 | 90.67 | 50.91 | 72.28 | 25.26 | 37.41 | 22.58 | 31.32 | 26.03 | 37.37 |

Table 1: Evaluation of four simple and three state-of-the-art models on three datasets. Merging synonyms largely boosts results in all cases, but there are inconsistencies between full- and zero-shot performance for most models. Notice how the simple image-agnostic baseline *LS* performs on par or better than state-of-the-art detectors, questioning the necessity of complex formulations, as well as scorning the "more than objects together" idea.

proving that the co-existence of multiple correct answers is indeed a major cause of confusion. The gain is consistent across the different baselines, approximately 24% for VRD and 23% for VG200 and VrR-VG (Table 1). However, spatial features (*S*) benefit more than 30% on the two VG splits, disproving previous works [62] that have doubted their importance; they are a strong cue if we circumvent the problematic annotations.

*Performance is not consistent on the zero-shot set.* On the full set, language is the most prominent feature and supplants visual features even on the "visually-relevant" VrR-VG, disclaiming the introduction of thousands of noisy object classes as a proper way to create an unbiased dataset. Nonetheless, things change radically on the zero-shot set, with language contributing the minimum and visual features the maximum to generalization.

These observations motivate a rethinking of the open margin to improve on these datasets. We find that many of the remaining errors are also due to synonyms that are not merged, since they are never annotated together. In fact, after manually merging the most common geometric predicates (e.g. "behind", "near" etc.), we are able to mine an approximately 4% of remaining errors on VRD, that are mostly hard and uncommon examples, e.g. in Fig. 2d, all models misdetect the person on the road, while she is standing on the pavement next to the road. Solving such errors is out of the scope of this paper; we instead propose an effective method to learn synonyms during training and regularize the output space of relations so that less frequent predicates are learned jointly with their more frequent synonyms.

## 3 Local Context and Synonymy

A scene graph is constructed by connecting subject and object nodes with predicate edges. The local context of an edge $P$ is the subject $S$ and the object $O$ of the relationship $<S, P, O>$ and can reason on the possibility of a predicate, as well its probability given that another predicate also holds. We present our visual-spatial baselines and then contrive a context-aware classification module and two loss terms to efficiently exploit local context. All models are trained on standard VRD/VG annotations.

**Visual-spatial architectures**: Our visual-spatial baseline *VS* (Fig. 3) fuses subject, predicate and object appearance features [28, 60], together with binary object masks [6] and box-deltas [61, 63]. The output of this module is an encoded relationship feature vector $F$.

Language can now be employed as a feature [28, 62], attention mechanism [9, 49] or classifier [9, 67]. We implement attention to show that even such implicit use of language is able to bias the model. Our scheme attends the object areas with the respective word embed-

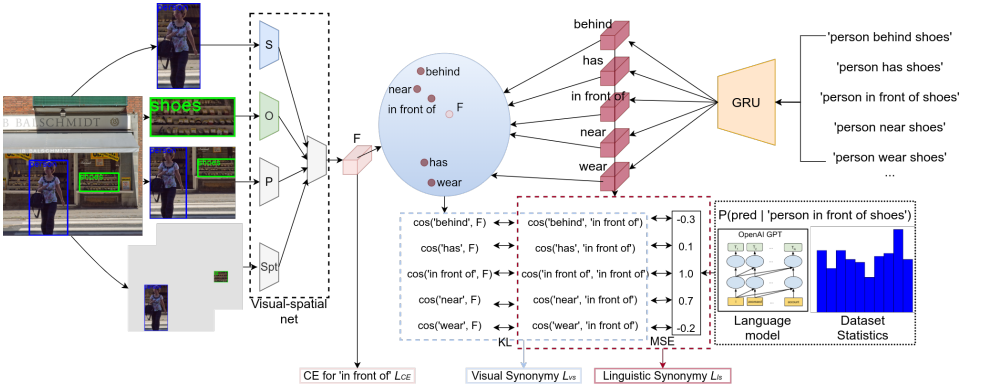


Figure 3: A network encodes appearance and location features into a vector $F$. A GRU encodes <$S,P,O$> phrases into local-context-aware classifiers. Features and classifiers are projected into a space where relationship scores are computed as cosine similarities. Both internal and external knowledge are exploited to pull classifiers of synonym predicates closer (linguistic synonymy) and boost the scores of the ground-truth's synonyms (visual synonymy).

dings [31] and the predicate area with the concatenation of subject and object embeddings. We add attentive and non- features element-wise and then fuse them with the spatial features as in the first variant. It is advantageous to deeply supervise the attentive areas, specifically, impose a predicate CE loss on the predicate area and object classification losses on the objects, forcing the attentive object features to represent the object classes, while non-attentive features represent the visual appearance. We experimentally prove (Sec. 4) that this model, *VSA* (visual-spatial-attention) tends to mimic *LS*, superseding other features.

**Context-aware classifiers** The above baselines learn a single classifier per predicate irrespective to its local context. We build local context-aware classifiers employing a bidirectional Gated Recurrent Unit (GRU) that encodes each relationship phrase <$S,P,O$>, with fixed $S$ and $O$, into a vector $R_P^{S,O}$. We then use this vector as a classifier: we project the network's features $F$ on it and measure their cosine similarity (Fig. 3). This is the score of the predicate $P$ for this pair of objects. We avoid adding bias to the output scores as this tends to favor the most frequent predicates. We concatenate these scores into a single vector, scale them by a factor $C$ and then apply the CE loss [52]. Thus, with $P_{gt}$ denoting the ground-truth predicate and $\sigma$ the softmax function, the classification loss:

$$L_{CE} = -log(\sigma(C * F * R_{P_{gt}}^{S,O})) \quad (1)$$

**Linguistic synonymy** Classifiers of the same predicate learn suitable templates under different local context. However, relying solely on the GRU to propose classifiers while supervising the output using a CE loss does not guarantee the semantic similarity of predicates. We design and apply a loss that transfers linguistic similarities to the classification space.

Annotating "ground-truth" similarities between predicates under different local context requires a significant and very costly effort if applied on scale on multiple datasets. Instead, our method is cheap and not bound to a specific dataset, based on the existing annotations and external knowledge. We first count the frequency of concurrent annotations for the same sample and then cluster relationships within a given local context w.r.t. spatial features to count co-occurrences in the same clusters. However, we have no synonymy estimation for

triplets that never appear on the dataset (∼99% of total). To tackle this, we use a language model [38] to estimate the probability that each predicate holds in the given context. We weight the internal and external ground-truth similarities into a single vector $M = \{M_{P,P_{gt}}\}$.

We compute a vector of cosine similarities between the classifier of the target predicate class $P_{gt}$ and every other classifier for this local context. We then force Mean Square Error (MSE) constraints between the computed and the ground-truth similarities (Fig. 3). Omitting superscripts $S, O$ for clarity, the linguistic synonymy ($ls$) loss obtains the form:

$$L_{ls} = \frac{1}{|P|}\sum_{P}(sim(R_P, R_{P_{gt}}) - M_{P,P_{gt}})^2 \tag{2}$$

where $sim$ the cosine similarity and $|P|$ the total number of predicates.

**Visual synonymy** Due to the extremely imbalanced training distribution, standard cross-entropy supervision tends to penalize the infrequent classes in favor of the most frequent. As we show in Sec. 4, a weighted or smoothed cross-entropy ignores the classes' semantics and turns out distractive. Instead, we propose a local-context-wise smoothing inspired by self-distillation approaches [64]. We force a Kullback-Leibler (KL) divergence loss between the output scores $s_P = \sigma(C * F * R^{S,O}_{P_{gt}})$ and the computed similarities $m_{P,P_{gt}} = sim(R_P, R_{P_{gt}})$, to obtain the visual synonymy ($vs$) loss:

$$L_{vs} = \sum_{P} m_{P,P_{gt}} log(m_{P,P_{gt}}) - \sum_{P} s_P log(m_{P,P_{gt}}) \tag{3}$$

The semantic similarities the classifiers share play the role of a teacher that distills knowledge about "what a classifier *could* confuse". Teacher and student are jointly trained, forcing the classifiers to adapt to both visual and semantic synonyms. Denoting as $L_{DS}$ the weighted sum of any deep supervision losses and $\lambda$ the balancing hyperparameters, the total objective is the weighted sum of the described entropy and synonymy losses:

$$L = \lambda_{CE} L_{CE} + \lambda_{ls} L_{ls} + \lambda_{vs} L_{vs} + \lambda_{DS} L_{DS} \tag{4}$$

**Implementation** We use Faster-RCNN [39] to extract features from the subject, object and predicate regions and each feature passes through a two-layer MLP. For spatial features we use the same net as in [6, 9, 63]. The concatenation of these four feature vectors is projected into a 128-dimensional classification space. The local-context-aware classifiers are constructed by vectorizing $<S, P, O>$ into word embeddings [31] and then feeding these sequences into the GRU. The whole network is trained end-to-end by optimizing Eq. 4 with the Adam optimizer [18]. We initiate the learning rate to 0.002 and multiply by 0.3 on validation loss plateaus. During the first few epochs, we use a larger $\lambda_{ls}$ so that the classifier learns early to propose synonyms, while we decrease its value over time. On the other hand, $\lambda_{vs}$ is increased over time, so that the model jointly classifies the synonyms proposed, instead of pursuing a single ground-truth ($\lambda_{CE}$). For further details and reproduction of our results, we make our code publicly available[1].

# 4 Experiments

Our aim is to increase zero-shot generalization with the minimum sacrifice of full set performance. We train our models on VRD but also test them on sVG [6] and UnRel [32]. We do

[1] https://github.com/deeplab-ai/zs-vrd-bmvc20

| Model | VRD | | VRD zs | | VRD to UnRel | | VRD to sVG | | |
|---|---|---|---|---|---|---|---|---|---|
| | top-1 | merged | top-1 | merged | top-1 | top-3 | top-1 | top-2 | top-3 |
| *V* | 52.59 | 77.00 | 24.57 | 36.47 | 16.17 | 30.4 | - | - | - |
| *L* | 54.28 | 79.55 | 17.12 | 28.77 | 10.13 | 15.17 | - | - | - |
| *S* | 47.58 | 70.49 | 25.51 | 36.82 | 10.67 | 29.33 | - | - | - |
| *LS* | 57.56 | 82.06 | 26.37 | 39.55 | 10.00 | 21.41 | - | - | - |
| *ATR-Net* | ***58.48*** | ***82.76*** | 27.10 | 40.50 | 10.67 | 24.70 | - | - | - |
| *UVTransE* | 57.25 | 81.50 | 27.48 | 40.58 | 15.77 | 31.74 | - | - | - |
| *RelDN* | 54.47 | 78.90 | 25.26 | 37.41 | 17.11 | 30.94 | - | - | - |
| *VS* | 54.55 | 78.79 | 26.63 | 39.38 | 17.52 | *34.23* | - | - | - |
| *VSA* | 56.68 | 81.43 | 26.11 | 39.13 | 14.36 | 26.44 | - | - | - |
| *VS-LoC* | 57.65 | 82.07 | *28.00* | *41.52* | *18.05* | 32.89 | ***52.76*** | *64.00* | *71.41* |
| *VS-LoC-ls-smth* | 50.81 | 78.17 | 24.23 | 36.39 | 12.95 | 26.24 | 38.51 | 56.28 | 65.34 |
| *VS-LoC-ls-wght* | 28.71 | 61.7 | 8.48 | 15.67 | 17.92 | 31.48 | 45.11 | 59.32 | 66.61 |
| *LS-LoC-ls-vs* | 57.02 | 81.77 | 25.34 | 38.7 | 14.90 | 24.70 | 41.52 | 54.05 | 62.57 |
| *VS-LoC-ls-vs* | 57.02 | 81.30 | 26.8 | 39.90 | ***18.12*** | ***34.68*** | *50.30* | ***64.78*** | ***72.43*** |
| *VSA-LoC-ls-vs* | *57.78* | *82.15* | ***28.77*** | ***42.20*** | 14.43 | 29.26 | 48.50 | 60.05 | 67.95 |

Table 2: Results of all examined models on the different setups used in our evaluation. *VS(A)-LoC-ls-vs*, the two visual-spatial(-attention) models employing local context, language synonymy and visual synonymy, show clear gain and state-of-the-art results on zero-shot tasks.

not train on VG, as testing on different partitions could blend already seen training images into the evaluation. While past approaches have adopted recall variants [3, 28, 48], accuracy [52] and mean average precision [32], as suitable metrics to their setup, we insist on accuracy as it assesses the recognition of each sample equally and separately. We present results for both original and merged annotations.

We first examine if fusion with language is beneficial for a visual-spatial network (*VS* versus *VSA*) and then we question the importance of the proposed components and losses: *LoC* denotes the local-context-aware classifier, *ls* the linguistic and *vs* the visual synonymy loss. We compare *vs* with other smoothed (*smth*) [40] and weighted (*wght*) [5] cross-entropy. The results for both baselines and ablative models are summarized in Table 2, where the advantages of the less biased visual-spatial models are obvious, especially on UnRel/sVG.

**Ablation on VRD/VRD-zs:** While *VS* performs far worse than *LS* on VRD full set, they perform on par on zero-shot. *VSA* largely improves over *VS*, with a slight decrease though on zero-shot accuracy: language pushes *VSA* to classify, and get biased by, the objects' labels.

Adding *LoC* boosts *VS*' performance by an absolute 3.1%, on par with with *LS* on the full set, but with more than 1.5% higher zero-shot accuracy. Intuitively, local-context-aware classifiers adapt to the objects' categories and can alleviate the high intra-class variance of predicates in different context, but, they rely on language and therefore suffer from the bias problem. The two synonymy losses are used to regularize the use of language on the classifier level; *VS-LoC-ls-vs* performs worse than *VS-LoC*, but, as we discuss later, it outperforms other ablative models on the harder evaluation tasks. Interestingly, *VSA-LoC-ls-vs* only slightly improves over *VS-LoC-ls-vs*, despite the margin between *VS* and *VSA*, further validating that significant portion of *LoC*'s is due to language. Still, *VSA-LoC-ls-vs* achieves a new state-of-the-art on zero-shot VRD over *UVTransE* and a second-best on the full set.

Table 2 also indicates that entropy weighting *VS-LoC-ls-wght* fails on VRD. The examples per class on VRD range from over 7000 to fewer than 10; even worse, rare classes are mostly synonyms of more frequent classes. Thus, penalizing misclassification of the tail forces the classifier to drastically (and incorrectly) adapt its weights to avoid predicting a

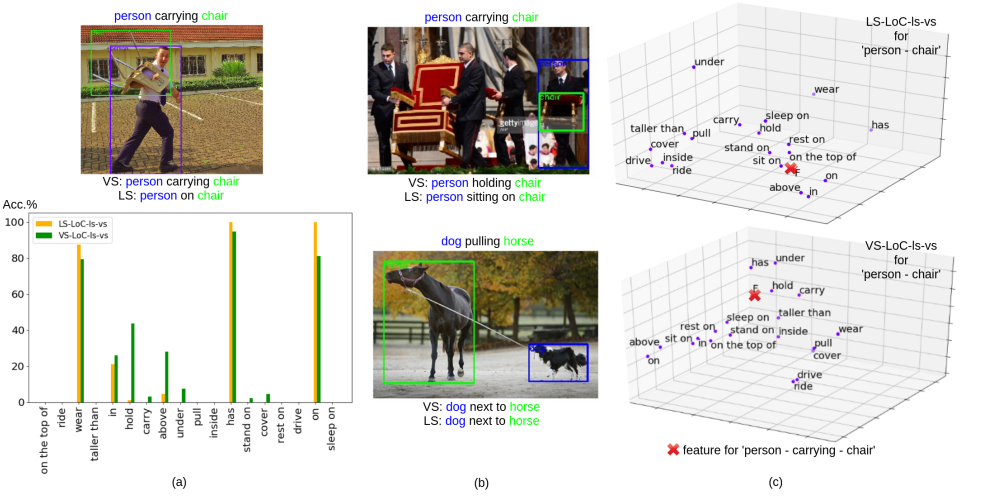


Figure 4: Results on UnRel: (a) *VS-LoC-ls-vs* is able to detect rare predicates, while *LS-LoC-ls-vs* perfectly predicts the most frequent classes "on" and "has". (b) *VS-LoC-ls-vs*'s predictions are still relevant in failure cases. (c) Classification spaces created by *LS-LoC-ls-vs* (up) and *VS-LoC-ls-vs* (down) for 'person - chair' and projected feature vector $F$ (red-cross) for the respective failure case of (b). Biased by linguistic features, *LS-LoC-ls-vs* projects $F$ near "sit on", while *VS-LoC-ls-vs* projects it on the correct synonym cluster ("hold", "carry").

frequent synonym. On the other hand, *VS-LoC-ls-smth* distills information about all classes uniformly and it is inferior to *vs* that learns to propose synonyms.

Lastly, our analysis proves that VRD zero-shot annotations also suffer from problems related to synonyms and that proper learning of a rare predicate does not mean it will be selected over a much more frequent synonym. Instead, our models tend to predict frequent visually-grounded predicates. As qualitatively explained (Fig. 4), the on par with state-of-the-art performance of the image-agnostic baseline is due to memorizing the dataset's bias, making its predictions irrelevant for data drawn from different distributions. In contrast, our models tend to predict the correct cluster of synonyms, even if they do not rank the expected ground-truth predicate above other valid ones.

**Zero-shot evaluation on UnRel:** UnRel contains of samples of very unusual triplets that never appear on VRD and cannot be inferred using common-sense. *VS-LoC-ls-vs* clearly outperforms all other variants, including the previous state-of-the-art *UVTransE* [14], stretching the importance of regularizing using the synonymy losses. Note the latent inverse relation between results on VRD and UnRel: *RelDN* [63] and *VS-LoC-ls-wght*, although significantly outperformed by other approaches on VRD, are able to generalize on UnRel. In contrast, the zero-shot transferability of *VSA-LoC-ls-vs* is not consistent on UnRel (Table 2).

We qualitatively compare *VS-LoC-ls-vs* to the improved *LS* baseline *LS-LoC-ls-vs* in Fig. 4a. Not only it achieves fairly higher accuracy, but even for classes that both models fail to predict, *VS-LoC-ls-vs*'s predictions are still relevant (Fig. 4b). For example, we counted how many times a model confuses "person holding chair" with "person sitting on chair", two non-synonym relations, to find an error rate of approximately 14% for *VS-LoC-ls-vs* versus 76% for *LS-LoC-ls-vs*, proving that *VS-LoC-ls-vs* can more efficiently learn synonyms and mitigate confusion between non-synonyms (Fig. 4c). Accuracy is still low since VRD and

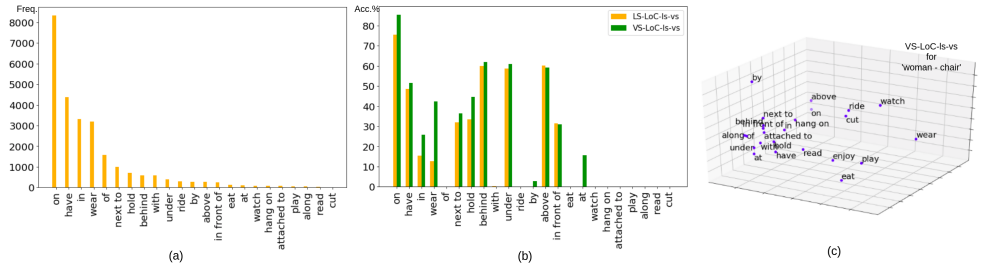


Figure 5: (a) The test distribution of sVG is similar to that of VRD, yet the classes are different. (b) *LS-LoC-ls-vs* mostly detects samples of classes respecting the VRD biases ("on", "next to", "behind"), while *VS-LoC-ls-vs* predicts both more classes and with higher accuracy. (c) *VS-LoC-ls-vs*'s classification space respects the semantic similarities of predicates (e.g."on" lies closest to "hang on" and "above", while "with", "hold" and "have" are neighbors), despite "woman" not being in the vocabulary of VRD.

UnRel have a different test distribution. In fact, most failure cases regard predicate synonyms that appear more times in the training set.

**Zero-shot transfer on sVG:** Our context-aware classifiers are not restricted on specific classes; we can insert new class names in the GRU to estimate a classifier on demand. We test on sVG models trained on VRD, despite the different classes and distribution, benchmarking a novel large-scale zero-shot scenario that extends from unseen triplets to novel object and predicate classes as well; from 399 objects on sVG, only 90 exist on VRD.

As seen in Table 2, *VS-LoC-ls-vs* vastly outperforms *VSA-LoC-ls-vs* and *LS-LoC-ls-vs* and improves over *VS-LoC* on higher accuracy thresholds. A qualitative analysis is also included in Fig. 5, as well as a comparison to the image-agnostic baseline, where it is obvious that the visual-spatial model can detect a wider variety of predicates. The supervised state-of-the-art for sVG is $\sim$77% [9], indicating a significant gap, but our error-analysis attributes many errors to VRD using "on" to describe "partOf" relations for which sVG uses "of".

## 5 Related Work

**Visual Relationship Detection** [28] is usually approached as a two-step task of first detecting objects and then classifying the predicate of each pair [9, 14, 25, 28, 54, 58, 62, 63, 67]. On a similar fashion, **Scene Graph Generation** [48] works, extract object proposals and jointly classify relationships on a graph [2, 3, 9, 10, 24, 29, 37, 41, 44, 47, 48, 51, 57]. Even when assuming a perfect object detector, such methods struggle to outperform an image-agnostic baseline knowing only of the objects' names and locations. Our work steps further on predicate classification, aiming to demystify the saturation of the results.

**Local context**: Conditioning on object classes has been widely adopted by previous works [3, 6, 9, 14, 25, 28, 34, 36, 55, 57, 58, 63, 65, 67]. While [9, 19, 67] construct local-context-aware classifiers, they still learn a fixed number of predicates, contrary to our recurrent classifiers that are created on demand for an infinite number of classes, allowing training and testing on different datasets. On the other hand, [62] employs similarity learning to handle a large, possibly open vocabulary, but ignores local context. The closest to ours are [23, 33, 35] that use local-context to project relationship embeddings in a multimodal space.

Nonetheless, they apply no similarity constraints to explicitly model synonym predicates.

**Zero/Few-Shot** detection [28] has emerged due to the quadratic number of relationships with respect to the examined objects. While prior methods often provide such results [25, 26, 28, 52, 55, 58, 60, 66, 67], only a few explicitly examine zero- or few-shot solutions [4, 7, 14, 32, 33, 45]. We extend the common zero-shot setup from unseen triplets to totally unseen object and predicates via testing on a different dataset and distribution. A concurrent stream of works attribute the long-tail distribution to "biased" annotations and attempt to increase recall of tail-classes [1, 3, 42, 46]. These approaches, however, do not explain the cause of such biases, i.e. predicate synonymy, instead they tend to overfit introduced metrics, resulting to a large drop on standard recall/accuracy metrics [56]. Lastly, [11] also identifies the problem of co-existing predicates, but does not propose a learnable solution.

**Unbalanced classification**: Our synonymy losses draw inspiration from different areas. First, to obstruct uncontrollable increase for more frequent classes, [14, 17] penalize classifiers' weights' norms; instead, we apply cosine *similarity learning* between generated classifiers and projected features [23, 33, 35, 62]. Next, *class weighting* [5, 56] and *smoothing* [40] are commonly applied to mitigate overfitting to most frequent classes, ignoring, however, predicate synonymy; in contrary, our proposed loss uses synonyms to smooth entropy. Lastly, visual synonymy can be seen as a *teacher-student imitation* loss [12, 34, 53], where the teacher proposes alternative predicates to describe a relation using internal and *external knowledge* [10, 25, 34, 55, 58]. Related approaches rely on not local-context-aware rule distillation [13], while our scheme is more reminiscent of self-distillation [64], using learned language synonymy to guide visual classification.

## 6 Conclusions

Penalizing co-existing synonym predicates renders visual relationship detection on existing datasets an ill-posed and non-causal problem. Models that employ language on the feature level are cursed to mimic the dataset's distribution, leading to saturated performance. We propose local-context-aware classifiers and synonymy losses, to adaptively classify predicates under different object context while respecting the semantic similarity of predicates. Our losses can effectively regularize the output space and amplify zero-shot generalization on extreme setups, such as transferring predictions between different datasets, achieving state-of-the-art results with minimal decrease on full set performance. Lastly, our analysis showcases the remaining margin for improvement on current datasets, refocusing generation of structured image representations around a new direction.

## Acknowledgments

This project was conducted while N. Gkanatsios was an intern at the National Technical University of Athens and has been funded by deeplab.ai, as far as V. Pitsikalis and N. Gkanatsios are concerned. This is a part of the deeplab.ai research activities, such as student research-training funding, and collaborations with academic institutions.

# References


[1] Sherif Abdelkarim, Panos Achlioptas, Jiaji Huang, Boyang Li, Kenneth Ward Church, and Mohamed Elhoseiny. Long-tail Visual Relationship Recognition with a Visiolinguistic Hubless Loss. *ArXiv*, abs/2004.00436, 2020.

[2] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based Weakly-supervised Learning of Visual Relations with Graph Networks. In *Proc. ECCV*, 2020.

[3] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-Embedded Routing Network for Scene Graph Generation. In *Proc. CVPR*, 2019.

[4] Vincent S. Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Ré, and Li Fei-Fei. Scene Graph Prediction With Limited Labels. In *Proc. ICCV*, 2019.

[5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *Proc. CVPR*, 2019.

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

[7] Apoorva Dornadula, Austin Narcomey, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual Relationships as Functions: Enabling Few-Shot Scene Graph Prediction. In *Proc. ICCV Workshops*, 2019.

[8] Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind. Generating Natural Language Explanations for Visual Question Answering using Scene Graphs and Visual Attention. *ArXiv*, abs/1902.05715, 2019.

[9] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, and Petros Maragos. Attention-Translation-Relation Network for Scalable Scene Graph Generation. In *Proc. ICCV Workshops*, 2019.

[10] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene Graph Generation with External Knowledge and Image Reconstruction. In *Proc. CVPR*, 2019.

[11] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. Learning Canonical Representations for Scene Graph to Image Generation. In *Proc. ECCV*, 2020.

[12] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. *ArXiv*, abs/1503.02531, 2015.

[13] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing. Harnessing Deep Neural Networks with Logic Rules. *ArXiv*, abs/1603.06318, 2016.

[14] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Union Visual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. *ArXiv*, abs/1905.11624, 2019.

[15] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proc. CVPR*, 2015.

[16] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image Generation from Scene Graphs. In *Proc. CVPR*, 2018.

[17] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In *Proc. ICLR*, 2019.

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

[19] Rajat Koner, Poulami Sinhamahapatra, and Volker Tresp. Relation Transformer Network. *ArXiv*, abs/2004.06193, 2020.

[20] 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*, 123, 2016.

[21] Ranjay Krishna, Ines Chami, Michael S. Bernstein, and Li Fei-Fei. Referring Relationships. *CoRR*, 2018.

[22] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. Learning Visual Relation Priors for Image-Text Matching and Image Captioning with Neural Scene Graph Generators. *ArXiv*, abs/1909.09953, 2019.

[23] Binglin Li. Visual Relationship Detection Using Joint Visual-Semantic Embedding. In *Proc. ICPR*, 2018.

[24] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene Graph Generation from Objects, Phrases and Region Captions. In *Proc. ICCV*, 2017.

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

[26] Xiaodan Liang, Lisa Lee, and Eric P. Xing. Deep Variation-Structured Reinforcement Learning for Visual Relationship and Attribute Detection. In *Proc. CVPR*, 2017.

[27] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. VrR-VG: Refocusing Visually-Relevant Relationships. In *Proc. ICCV*, 2019.

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

[29] Li Mi and Zhenzhong Chen. Hierarchical Graph Attention Network for Visual Relationship Detection. In *Proc. CVPR*, 2020.

[30] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. Interactive Image Generation Using Scene Graphs. In *Proc. ICLR*, 2019.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. In *Proc. EMNLP*, 2014.

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

[33] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Detecting Unseen Visual Relations Using Analogies. In *Proc. ICCV*, 2019.

[34] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Françoise J. Prêteux. Visual Relationship Detection Based on Guided Proposals and Semantic Knowledge Distillation. In *Proc. ICME*, 2018.

[35] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Françoise J. Prêteux. Learning Prototypes for Visual Relationship Detection. In *Proc. CBMI*, 2018.

[36] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In *Proc. ICCV*, 2017.

[37] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive Relational Networks for Mapping Images to Scene Graphs. In *Proc. CVPR*, 2019.

[38] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

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

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proc. CVPR*, 2016.

[41] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Weiwei Liu. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *Proc. CVPR*, 2019.

[42] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased Scene Graph Generation from Biased Training. In *Proc. CVPR*, 2020.

[43] Subarna Tripathi, Anahita Bhiwandiwalla, Alexei Bastidas, and Hanlin Tang. Using Scene Graph Context to Improve Image Generation. *ArXiv*, abs/1901.03762, 2019.

[44] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring Context and Visual Pattern of Relationship for Scene Graph Generation. In *Proc. CVPR*, 2019.

[45] Xiaogang Wang, Qianru Sun, Tat-Seng Chua, and Marcelo H. Ang. Generating Expensive Relationship Features from Cheap Objects. In *Proc. BMVC*, 2019.

[46] Bin Wen, Jie Luo, Xianglong Liu, and Lei Huang. Unbiased Scene Graph Generation via Rich and Fair Semantic Extraction. *ArXiv*, abs/2002.00176, 2020.

[47] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In-So Kweon. LinkNet: Relational Embedding for Scene Graph. In *Proc. NeurIPS*, 2018.

[48] Danfei Xu, Yuke Zhu, Christopher Bongsoo Choy, and Li Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *Proc. CVPR*, pages 3097–3106, 2017.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2015.

[50] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. Scene Graph Captioner: Image Captioning Based on Structural Visual Representation. *Visual Communication and Image Representation*, 2019.

[51] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for Scene Graph Generation. In *Proc. ECCV*, 2018.

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

[53] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring Visual Relationship for Image Captioning. In *Proc. ECCV*, 2018.

[54] 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.

[55] 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.

[56] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging Knowledge Graphs to Generate Scene Graphs. In *Proc. ECCV*, 2020.

[57] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene Graph Parsing with Global Context. In *Proc. CVPR*, 2018.

[58] Yibing Zhan, Jia Ming Yu, Ting Yu, and Dacheng Tao. On Exploring Undetermined Relationships for Visual Relationship Detection. In *Proc. CVPR*, 2019.

[59] Cuicui Zhang, Wei-Lun Chao, and Dong Xuan. An Empirical Study on Leveraging Scene Graphs for Visual Question Answering. In *Proc. BMVC*, 2019.

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

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

[62] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed M. Elgammal, and Mohamed Elhoseiny. Large-Scale Visual Relationship Understanding. In *Proc. AAAI*, 2019.

[63] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical Contrastive Losses for Scene Graph Generation. In *Proc. CVPR*, 2019.

[64] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In *Proc. ICCV*, 2019.

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

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

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