Generating Expensive Relationship Features from Cheap Objects

Xiaogang Wang[1]
xiaogangw@u.nus.edu
Qianru Sun[2,3*]
qsun@mpi-inf.mpg.de
Tat-Seng Chua[1]
chuats@comp.nus.edu.sg
Marcelo H Ang Jr[1]
mpeangh@nus.edu.sg

[1] National University of Singapore
[2] Singapore Management University
[3] Max-Planck Institute for Informatics

Abstract

We investigate the problem of object relationship classification of visual scenes. For a relationship *object1-predicate-object2* that captures the object interaction, its representation is composed by the combination of object1 and object2 features. As a result, relationship classification models usually bias to the frequent objects, leading to poor generalization to rare or unseen objects. Inspired by the data augmentation methods, we propose a novel Semantic Transform Generative Adversarial Network (ST-GAN) that synthesizes relationship features for rare objects, conditioned on the features from random instances of the objects. Specifically, ST-GAN essentially offers a semantic transform function from cheap object features to expensive relationship features. Here, "cheap" means any easy-to-collect object which possesses an original but undesired relationship attribute, e.g., *a sitting person*; "expensive" means a target relationship on this object, e.g., *person-riding-horse*. By generating massive triplet combinations from any object pair with larger variance, ST-GAN can reduce the data bias. Extensive experiments on two benchmarks – Visual Relationship Detection (VRD) and Visual Genome (VG), show that using our synthesized features for data augmentation, the relationship classification model can be consistently improved in various settings such as zero-shot and low-shot.

1 Introduction

We study the task of visual relationship classification that labels an image region by a triplet composed of object1, predicate and object2. Visual relationship classification is essentially a compositional recognition and its representation is extracted from the image regions labeled as object1 and object2. Taking *person-riding-horse* in Figure 1 as an example, *riding* is represented by the attributes of *person* and *horse*. It is hard to annotate an accurate image region for the predicate *riding* itself. Existing works [25, 60, 63, 64, 67] proposed deep networks to pass messages between object1 and object2, to learn a classifier on different kinds of relationships. Their methods thus suffer from the problem that the recognition of

* Corresponding Author.

visual relationships is dominated by the distribution of objects. An evidence was reported by Zellers *et al.* [66] that the relationship recognition models tend to be seriously skewed once the categories of object1 and object2 are known. Such objective bias results in the poor generalization ability of the model in recognizing rare or unseen object combinations [63], *e.g.*, on Visual Genome (VG) dataset [21], the recognition rate of unseen classes is only 18.9%, *i.e.*, 44.0% lower than the average of all classes.

In this paper, we tackle this problem by augmenting relationship features for rare or unseen categories. Existing data augmentation methods [17, 33, 49, 59] focus on single image or object feature synthesizing, which cannot be directly applied to our task, because our relation representation has triple components *object1, predicate, object2*, which requires preserving semantic information consistent with image context. Therefore, we propose a novel relationship feature generation model – Semantic Transform Generative Adversarial Networks (ST-GAN). Here, "semantic" feature means the higher-level network layer output such as the feature at the last layers of VGG17 [50] and ResNets [18]. These semantics are "transformed" from the original undesired relationships of objects (on input images) to a target relationship of the same object pair. Taking Figure 1 as an example, "cheap" means any easy-to-collect objects – *person* and *horse*, which possess their original but undesired relationships *jumping* and *pulling a cart* on random images (see Figure 1). "Expensive" means a target compositional relationship of these two objects – *person-riding-horse*. More specifically, this means the annotation process of *person-riding-horse* is more expensive than that of the object *person* (or *horse*).

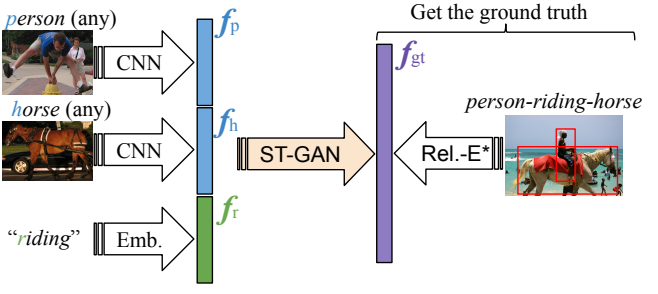


Figure 1: Our ST-GAN synthesizes visual relationship feature given a triplet label: *obj1-pred-obj2*, *e.g. person-riding-horse*. The inputs for ST-GAN are: 1) CNN features of *obj1* and *obj2* from any same-labeled instances, *i.e.cheap* objects; 2) the word vector of *pred* as the reference of the target relationship. The feature extracted from the triplet *obj1-pred-obj2* serves as the ground truth. Rel.-E* indicates a relationship feature extractor, *e.g.* Motifs [66] and VTransE [67]. Emb. is an embedding model, *e.g.* GloVe [41] and word2vec [36].

With ST-GAN, we aim to transfer the shared objective variance from many-shot relations to the low-shot or zero-shot relations. In order to validate the usefulness of ST-GAN features, we utilize VTransE [67] as our baseline model. This model represents object relations purely by the object appearance features without any knowledge from external corpus [63] or complicated network architectures [25, 64]. For handling zero-shot objects, we leverage word embeddings [36] as generation conditions for desired relationship categories, inspired by [59]. Our model is trained on the whole dataset, then synthesize features for desired relationship categories.

Our main contributions are in three folds. (1) We introduce a novel feature generation model ST-GAN that synthesizes expensive relationship features from cheap object features. We demonstrate that ST-GAN is particularly effective for augmenting features of zero-shot and low-shot categories and improving their recognition performance by a large margin. (2)

We observe consistent improvements brought by ST-GAN over ablative baselines and the state-of-the-art methods on the VRD [31] and VG [21] datasets. (3) We validate the framework of ST-GAN is generalizable by applying different architectures as well as different kinds of reference information for the target relationship.

2 Related work

Visual relationship detection. Related methods for visual relationship detection can be roughly categorized into two categories. (1) Generic visual relationship detection methods. Some works [19, 23, 25, 53, 58, 60, 61, 62, 64, 67] focus on predicting the relationship by passing messages between objects or considering them as a whole. [67] proposes a simple yet effective method by concatenating objects features directly, then these features are used to train one relationship recognition model. Their features are object-level and thus can be easily used as the training data for ST-GAN. [63] aims to remove the data bias by shuffling and then assembling object categories in a triplet, thus forcing the relationship features to be object-agnostic. Interestingly, on the contrary to [63], another category of methods make use of such objective statistics (bias). (2) Subject-object statistics based methods. [6, 66, 69] mine the statistics on the dataset bias, then the motif patterns are used for relationship detection, and obtain impressive results. Some other works [4, 8, 15, 43] utilize external knowledge to obtain the scene graph, but none of them can handle the class-imbalance problem.

Few-shot learning. We divide few-shot learning methods into three categories. 1) Metric learning methods [45, 48, 51, 57] learn an metric embedding space by a large corpus of know categories instances, then some metric (L_1 or L_2) is used to classify instances of new categories by the proximity to few labeled training samples. These approaches learn a quite meaningful semantic space and show great performance. 2) Meta-learning methods [11, 14, 22, 44, 55] learn few-shot tasks instead of specific object instances, hence the learned models can learn new tasks with few labeled data. E.g. MAML [11], Meta-SGD [26], MTL [54] and LCC [30] have their models effectively learned with small data by meta gradient descent. 3) Generative methods [17, 34, 37, 49] learn to synthesize new data based on few training examples, or additional samples by some transfer learning [5, 20, 32, 56] from external data. For example, [17] present an approach of synthesizing additional examples for the data-starved classes, meanwhile. But for augmentation, they need at least two pairs of same category data and cannot be applied to zero-shot or even one-shot problems. [49] propose a modified auto-encoder to synthesize new samples by transferring the learned variances to unseen classes.

Zero-shot learning. In this setting, test classes are unseen during training [2, 10, 29, 52, 59]. We divide zero-shot learning methods into three categories. 1) [39, 70] regard the unseen classes as a mixture of seen class proportions; 2) [12, 46] make use of transductive learning to recognize unseen classes; 3) GAN [13] based models [3, 17, 33] generate additional samples in feature space for zero-shot classes. They aim to tackle the object classification problems, but not directly applicable for fine-grained triplet classes, i.e. relations.

Word embedding. Word embedding is widely studied in natural language processing (NLP) [7] related tasks. It has been shown a superior performance for measuring similarities and dissimilarities among words. This has given rise to many word representation model such as Skip Gram and Common Bag Of Words (CBOW) [35]. In this paper, we leverage the word representation as the condition of feature generation for desired relationship categories.

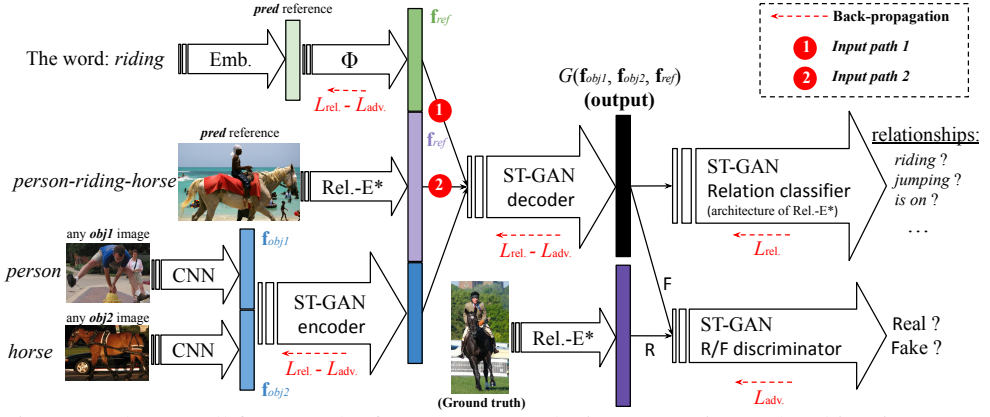


Figure 2: The overall framework of our ST-GAN. The input contains 1) the object instances from random images containing *cheap* objects, and 2) one kind of references of target relation, i.e., the label or the image example containing this relation. The generation is penalized by the real/fake discrimination loss and the relation classification loss (the $L1$ loss between the generated feature and the ground truth feature is used, see Section 3.4).

3 Semantic Transform Generative Adversarial Networks

Our ST-GAN aims to transform the original relationship attributes of *cheap* objects for generating *expensive* features of the target relationship. Its overall framework, shown in Figure 2, includes four main components namely ST-GAN encoder, ST-GAN decoder, ST-GAN R/F (real/fake) discriminator and ST-GAN relation classifier. Following related augmentation methods [9, 17, 49, 59], we first pre-train the word and image embedding modules, i.e., Emb., CNN and Rel-E*, and then fix them.

As demonstrated in Figure 2, for a target triplet (*obj1*, *pred*, *obj2*), we first sample *obj1* and *obj2* instances from random images. Then, we feed their CNN features to ST-GAN encoder (Section 3.1). The encoded feature (deep blue bar) together combined with the reference information of the target relation go to ST-GAN decoder for generation (Section 3.2). Note that *Input path 1* and *Input path 2* show two options of the reference information – the word embedding of relation label (deep green bar) *OR* an image feature example of the target relation (light purple bar). Note that only the word embedding works for the zero-shot case. Finally, ST-GAN components are optimized by the losses of R/F discriminator and relation classifier, as well as the $L1$ distance to the ground truth feature (extracted from a random sample of the target relation), see Section 3.4.

3.1 ST-GAN encoder *Enc*

ST-GAN encoder *Enc* is a function to transform the CNN features of new object instances to an intermediate feature $\mathbf{z}$ which is expected to contain the object category-related features without original relationship attributes (from source images).

Input: *cheap* objects. For obj_1 and obj_2, we randomly sample new instances from a source database according to object labels and locations. Basically, we do not need any relationship annotation. This is why we call these new instances as *cheap* objects. We choose *cheap*

objects only from the dataset we are working on[1], e.g., VRD [31] and VG [21]. The relationship detection dataset has the object labels and locations, according to which we can pick up cheap objects directly. The relation annotation on source images are additional information we might use. Taking object obj_1 in relation (obj_1, $pred$, obj_2) as an example, a new instance of obj_1 can be chosen from an image with the relation label as (obj_1, $pred$, obj_x) *OR* with the relation label as (obj_1, $pred_x$, obj_x), for which x means "any". In experiments, we name these choices as *Ran*. (*pred* is random, less annotation) and *Sel*. (*pred* is selected), respectively. As shown on the left in Figure 2, new instances of obj_1 and obj_2 are first represented as CNN features $\mathbf{f}_{obj1}$ and $\mathbf{f}_{obj2}$, for which the CNN was trained on ImageNet [47]. Then, the concatenation of $\mathbf{f}_{obj1}$ and $\mathbf{f}_{obj2}$ is used as input fed into ST-GAN encoder. This encoder has multi-layer perceptrons (MLP) with a leaky ReLU non-linearity except for the last layer (more details can be found in the supplementary materials).

3.2 ST-GAN decoder *Dec*

ST-GAN decoder *Dec* is conditioned on both the output of encoder *Enc* and one kind of references of the target relationship (*pred*). It aims to embed the variance of new objects encoded in *Enc* output into this predicate, in order to generate a new relationship feature containing the semantics of both.

Choosing relationship (*pred*) reference $\mathbf{f}_{ref}$. We have two options for referencing to the target relationship, as shown in Figure 2. *Input path 1*: the word embedding vector, denoted as $\mathbf{f}_{ref}^{(w)}$, of the relationship label. This works well for the zero-shot setting. The vector is extracted from a trained language embedding model, e.g. GloVe [41] and word2vec [36], denoted as Emb. In order to reduce the gap between language and image feature spaces, we propose to use a trainable Φ function following the language model. *Input path 2*: the relation feature denoted as $\mathbf{f}_{ref}^{(s)}$ extracted from a ground truth example. The feature extraction model Rel-E* was trained (also fixed) in prior using existing methods, e.g., VTransE [69]. Note that this path is not working for the zero-shot case, as there is no ground truth image.

Choosing images for $\mathbf{f}_{ref}^{(s)}$. As $\mathrm{f}_{ref}^{(s)}$ is extracted from an image example, the relation label of this example can have the same-predicate or the exact same-relation (triplet) to the target. For example, when given the target relation (obj_1, $pred$, obj_2), the label of a selected reference image can be either (obj_x, $pred$, obj_x) or (obj_1, $pred$, obj_2). In general, same-relation images are more limited in real datasets, and it is particularly difficult to find enough examples for low-shot relation classes. On the other hand, we can find more same-predicate images using which may cause the conflicts of objective semantics between the reference and the target. In experiments, we use same-relation images by default but use same-predicate images for an ablation study.

3.3 ST-GAN discriminators

As shown in Figure 2, our ST-GAN contains two discriminators. The real/fake discriminator distinguishes between ground truth features (real) and generated features (fake). The relation classifier encourages the generated features to be useful for relationship classification.

Real/Fake discriminator. WGAN has been proved to be significantly effective for the generation of image features [17, 49, 59]. We extend the improved WGAN model – WGAN-

[1]Note that auxiliary data source is applicable in our paradigm assuming that 1) the data domain is not far away from our target relationship data, and 2) object labels and bounding boxes are given or easy to detect.

GP [16] to fit our situation of conditioning on two object images and one relation reference information. Specifically, we learn a conditional ST-GAN with the *Enc* taking the CNN features of $obj1$ and $obj2$, *i.e.*, $\mathbf{f}_{obj_1}$ and $\mathbf{f}_{obj_2}$, to produce an intermediate vector $\mathbf{z}$. Then, ST-GAN *Dec* takes $\mathbf{z}$ concatenated with a reference feature of the relation ($\mathbf{f}_{ref}^{(s)}$ or $\mathbf{f}_{ref}^{(w)}$) to generate the target feature as real as possible. To this end, we propose to optimize ST-GAN by the following WGAN-GP loss,

$$\mathcal{L}_{adv} = \mathbb{E}\big[D(x)\big] - \mathbb{E}\big[D(\tilde{x})\big] - \lambda \mathbb{E}\big[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2\big] \tag{1}$$

where $\hat{x} = \theta x + (1-\theta)\tilde{x}$ with $\theta \sim (0,1)$, and $\tilde{x}$ denotes the generated feature which is equal to $G[\mathbf{f}_{obj_1}, \mathbf{f}_{obj_2}, \mathbf{f}_{ref}]$. G represents the generator consisting three learnable components *Enc*, *Dec* and Φ, and D represents the real/fake discriminator. In this equation, the third term is the gradient penalty which enforces the gradient of discriminator to have a unit norm along the straight line between pairs of real and generated points and λ is the penalty coefficient [16].

Relation classifier. WGAN does not guarantee the generated feature is semantically close to the target one. Referring to related works [17, 40, 59], this issue could be alleviated by encouraging the model to optimize its parameters towards constructing features, that can be correctly categorized by standard relationship classifiers. We thus propose to minimize the relationship categorization loss over generated features. Specific formulation is as follows,

$$\mathcal{L}_{rel} = -\mathbb{E}\big[\log P(C(\mathbf{f}_{ref}) | G([\mathbf{f}_{obj_1}, \mathbf{f}_{obj_2}, \mathbf{f}_{ref}]); R)\big] \tag{2}$$

which is an empirical classification loss, e.g. cross-entropy loss. R represents the relation classifier module. Both G and R aim to minimize this objective. $C(\mathbf{f}_{ref})$ is the relationship label. It is the predicate (*pred*), e.g. *riding*, in the triplet e.g. *person-riding-horse*.

3.4 Full objective

We use $\mathbf{f}_{gt}$ to denote the ground truth feature. $L1$ loss is used to ensure that the generated feature is not far from the ground truth. Furthermore, WGAN-GP loss and the relationship prediction loss are jointly optimized. L_1 loss and the full objective are as follows,

$$\mathcal{L}_1 = \mathbb{E}\big[\|G([\mathbf{f}_{obj_1}, \mathbf{f}_{obj_2}, \mathbf{f}_{ref}]) - \mathbf{f}_{gt}\|_1\big]. \tag{3}$$

$$\min_G \max_D \alpha \mathcal{L}_{adv} + \beta \mathcal{L}_{rel} + \gamma \mathcal{L}_1 \tag{4}$$

where α, β and γ are hyperparameters that are manually set to weight the effects of three losses.

4 Experiments

We evaluate the proposed ST-GAN approach in terms of its performance for augmenting relationship features of all classes, low-shot classes, e.g., a class less than 5-shot in training data is categorized into "low-shot $n = 5$", and zero-shot test classes that never appear in training data. Note that we **train ST-GAN model using the whole dataset for once**, and then apply the trained model to generate features for all, low-shot and zero-shot classes (see detailed settings in Section 4.1). Finally, both the real and generated features are used to the train relationship recognition models.

4.1 Datasets and implementation details

Datasets. We use ST-GAN to generate relationship features and train the relation recognition model on two benchmarks: Visual Relationship Detection (VRD) [31] and Visual Genome (VG) [21]. For VRD, we follow the training/test split in works [31, 63, 67], *i.e.*, 4,000 images for training and 1,000 for test. For VG, we use preprocessed data and splits containing 73,794 images for training and 25,858 for test [63, 67].

Settings. Following the standard settings in related works [28, 31, 63, 67], we focus on the visual relationship classification which concerns the purified problem of visual relationship detection. This means we use ground truth bounding boxes to locate objects but do not consider noisy detected bounding boxes [23, 24, 38, 62, 66]. According to the different numbers of training data of different visual relation classes, we have following three evaluation settings. **All classes (ALL)** is the typical supervised setting that classify all relation classes. In this setting, we augment relation features to double the whole dataset. **Low-shot classes (LShot)** case focuses on the classification of low-shot relation classes. If the training sample number of a relation r is less than or equal to n, then the classification of r is in a n-shot classification regime. We have $n = 1, 5, 10, 20$. In this setting, we augment relation features only for low-shot classes. **Zero-shot classes (ZShot)** case only counts the classification accuracy of relations which appear in the test set but unseen in the training set.

Ablation settings. (1) Alternative architectures from related generation models [17, 49, 59]. In order to validate whether our framework is superior for generating useful additional data, we conduct same augmentation experiments using different data generation models; (2) ST-GAN without discriminator. This is an ablation test of adversarial training to verify the effectiveness of using adversarial training in our network, we switch off the GAN module by simply setting $\alpha = 0$; (3) ST-GAN without encoder. This setting is to verify that performance improvements are obtained by the ability of transferring information using ST-GAN encoder and decoder, as opposed to the sampling scheme for the inputs. More ablation studies can be found in supplementary materials.

Other details. We add one fully connected layer after the VGG model to extract the 500-dimension object feature, and the relation feature is the concatenation of two object features, *i.e.*1000-dimension. For ST-GAN, each training iteration consists of 1 and 5 updates for the generator (including Φ, *Enc* and *Dec*) and the discriminator (including Real/Fake discriminator and Relation classifier), respectively. As to generating new features: for ALL classes, we synthesize one new relation feature for each sample of the training data; for LShot and ZShot classes, we synthesize 5 and 20 times of new data, respectively. For the final evaluation, we use Recall@50 (**R**@50) and Recall@100 (**R**@100) as metrics. There are two kinds of calculation methods: one is with graph constraint and the other one is without graph constraint. Omitting graph constraint, namely, allowing a subject-object pair to have multiple predicate labels in system output. Our code is available at https://github.com/xiaogangw/Generating-Expensive-Relationship-Features-from-Cheap-Objects.git.

4.2 Results and analyses

In Table 1, we show the data augmentation results using our proposed ST-GAN, on the VRD and VG datasets. We then give the ablation study results in Table 2, and comparison to the state-of-the-arts in Table 3. Finally, in Figure 3, we demonstrate some success and failure cases of relationship classification, comparing to baseline models.

		VRD						VG					
		VTransE[67]	$\mathbf{f}^{(s)}_{ref}$		$\mathbf{f}^{(w)}_{ref}$		↑	VTransE[67]	$\mathbf{f}^{(s)}_{ref}$		$\mathbf{f}^{(w)}_{ref}$		↑
			Ran	*Sel*	*Ran*	*Sel*			*Ran*	*Sel*	*Ran*	*Sel*	
ZShot	*n*=0	18.4*	–	–	21.4	**21.7**	3.3	16.4*	–	–	18.0	**19.0**	2.6
LShot	*n*=1	20.9*	24.2	**24.4**	23.3	23.4	3.5	19.5*	20.4	**20.9**	20.1	20.4	1.4
	n=5	24.1*	27.1	**27.4**	26.1	26.2	3.3	22.9*	24.0	**24.0**	23.7	23.7	1.1
	n=10	27.4*	30.3	**30.6**	29.1	30.0	3.2	25.6*	26.7	**27.1**	26.2	26.3	1.5
	n=20	32.0*	34.5	**34.8**	33.5	33.9	2.8	28.1*	29.6	**30.0**	28.5	28.8	1.9
ALL	@50	44.8 (49.0*)	51.8	**52.0**	50.7	50.9	3.0	62.6	63.5	**63.7**	63.0	63.1	1.1
	@100	44.8 (49.0*)	51.8	**52.0**	50.7	50.9	3.0	62.9	63.9	**64.0**	63.4	63.5	1.1

Table 1: Relation recognition accuracy (%) using our feature augmentation method, comparing to a baseline model VtransE [67]. * indicates our implementation using their code.

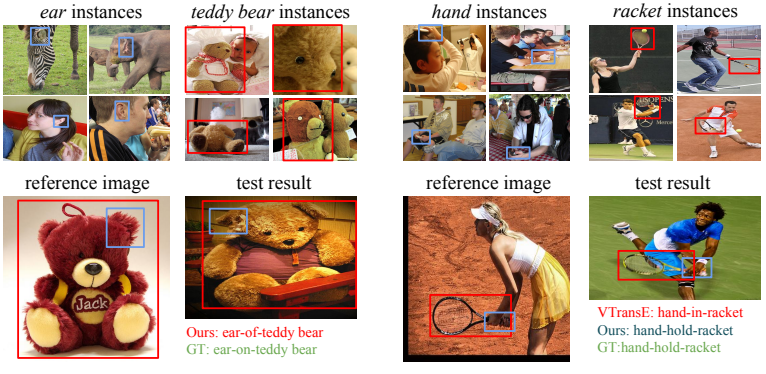
Augmentation results by ST-GAN. In Table 1, we present the results of data augmentation using the features generated by our ST-GAN. Real data and generated data are merged together to train relation recognition models. We take VTransE [67] as our baseline. Our generation model is trained with the ground truth extracted from a trained VTransE model. Using ST-GAN (trained only on the whole dataset), we can generate new features for low-shot classes (LShot), zero-shot classes (ZShot) and all classes (ALL). The maximum improvements are given in the columns denoted as ↑. Note that *Ran*. or *Sel*. is the input for the encoder and $f^{(s)}_{ref}$ or $f^{(w)}_{ref}$ is the input for the decoder. The notation *Ran*. means we randomly sample object instances according to object labels. *Sel*. means we sample object instances according to not only the object labels but also the labels of relationships the objects belonging to, i.e. these relationships should be same to the target one.

In Table 1, we have three observations. **(1) Fewer-data cases gain more improvements.** Compared to the case of ALL classes, LShot and ZShot cases gain larger improvements using our method. This fits well to the common sense that data augmentation methods bring more gains to the training on fewer data with the extreme case shown in zero-shot feature augmentation [59]. In terms of the datasets, we can see that VRD "enjoys more dividends" brought by our augmentation method, since VRD has less data than VG. **(2) Image reference is better than word reference.** On both datasets, we get more improvements using the image feature of the relation example ($\mathbf{f}^{(s)}_{ref}$) as reference than using the word embedding of relationship label ($\mathbf{f}^{(w)}_{ref}$), *e.g.*, about a 1% margin (for both *Ran* and *Sel*.) on the VRD dataset. The max improvements of most settings are obtained using $\mathbf{f}^{(s)}_{ref}$. However, this feature is not available for zero-shot classes which do not have any training sample. We use only the embedding feature $\mathbf{f}^{(w)}_{ref}$ in this zero-shot case, and obtain 3.3% and 2.4% improvements on two datasets, respectively. **(3) Random object images bring satisfied improvements with low annotation cost.** As mentioned, although *Sel*. have higher similarity to the target, we can see using *Sel*. is only slightly better than using *Ran*., e.g., 0.2% higher on VRD ALL. Therefore, using *Ran*. instances is a *cheaper* way to satisfy the augmentation requirement without losing much accuracy. This also validates that our ST-GAN can make successfully semantic transformations even for the instances with very different attributes.

Ablative study. We show the ablative study results in Table 2. We have following observations. **(1) Our method achieves the best performance** over other feature generation models [17, 49, 59]. Delta-encoder [49] utilized a simple reconstruction loss without any relation-specific penalty or adversarial loss. Even though the model [59] used two discriminator losses as ours, it still gets inferior performance. The possible reason is that it only

		VRD						VG					
		Other models			**Ours**			Other models			**Ours**		
		[42]	[63]	[67]	w/o *Dis*	w/o *Enc*	ST-GAN	[42]	[63]	[67]	w/o *Dis*	w/o *Enc*	ST-GAN
ZShot	n=0	–	20.2	19.4	20.0	20.3	**21.7**	–	17.2	16.7	17.4	17.1	**19.0**
LShot	n=1	21.1	21.6	21.8	22.7	23.0	**24.4**	19.6	19.7	20.2	19.9	19.7	**20.9**
	n=5	24.1	24.5	24.8	26.6	25.7	**27.4**	22.9	23.3	23.3	22.9	23.0	**24.0**
	n=10	27.3	27.5	27.8	30.0	30.2	**30.6**	25.5	25.7	25.6	26.0	25.8	**27.1**
	n=20	32.3	32.1	32.5	34.0	33.9	**34.8**	28.2	28.2	29.1	28.5	28.6	**30.0**
ALL	@50	48.8	49.1	49.6	51.3	51.2	**52.0**	62.4	62.7	62.6	63.0	63.1	**63.7**
	@100	48.8	49.1	49.6	51.3	51.2	**52.0**	62.6	63.0	63.0	63.3	63.4	**64.0**

Table 2: Ablation study results. Note that $\mathbf{f}_{ref}^{(s)}$ is used as a reference information by default.



(a) failure case (b) success case

Figure 3: (a) Our failure case. (b) Our success case. GT is ground truth. In (b) we also give the wrong predication by the baseline model – VTransE [67].

takes word embedding vector as input. While, our ST-GAN has more informative inputs both for the encoder and decoder, i.e., the image features extracted from *cheap* objects and the relation feature extracted from a reference image. **(2) ST-GAN R/F discriminator is helpful.** Inferior results are obtained when removing ST-GAN R/F discriminator. Since reconstruction loss alone can only guarantee the feature quality, however, learning good feature distribution is critical for generating new samples, and this can be achieved by adversarial training [1]. **(3) ST-GAN encoder is helpful.** We obtain lower accuracies when removing ST-GAN encoder and feeding reference feature directly into the decoder. This is because the encoder works for filtering out the object attributes related to the original relationships they belonging to [49].

Comparing to the state-of-the-art. In Table 3, we list the comparable[2] methods and results on VRD and VG datasets, respectively. On the VG dataset, there are different training/test splits and and we use the split-2 version following the most related works [63, 67]. We can see our method makes further improvements on ALL classes, and obtains larger margin under no graph constraints. This verifies our method on generating reasonable semantic relation features. Removing graph constraints significantly increases reported performance since the model is then allowed multiple guesses for challenging objects and relations, especially for zero-shot cases.

Success and failure cases. In Figure 3, we present the representative cases using source images: (a) the failure sample of our method, and (b) the success case which are wrongly

[2]Others [24, 65, 71] using strong prior knowledge are not directly comparable to our method.

	Method	VRD ALL @50	VRD ALL @100	VRD ZShot @50	VRD ZShot @100
GC	Motifs [56]	48.9*	48.9*	-	-
GC	VTransE [67]	49.0*	49.0*	18.4	18.4
GC	STA [53]	48.0	48.0	20.6	20.6
GC	**ST-GAN-f$^{(w)}_{ref}$**	50.9	50.9	**21.7**	**21.7**
GC	**ST-GAN-f$^{(s)}_{ref}$**	**52.0**	**52.0**	-	-
No GC	LVR [42]	46.3	47.9	8.5	8.5
No GC	PPR-FCN [65]	47.4	47.4	-	-
No GC	VDL [31]	47.9	55.16	19.17	19.17
No GC	DRL [6]	80.78	81.90	-	-
No GC	DSR [27]	60.90	79.81	-	-
No GC	**ST-GAN-f$^{(w)}_{ref}$**	**86.38**	**93.92**	**64.64**	**81.85**

		Method	VG ALL @50	VG ALL @100	VG ZShot @50	VG ZShot @100
GC	Split-1	SMP [50]	44.8	53.1	-	-
GC	Split-1	PAE [33]	54.2	55.5	-	-
GC	Split-1	Motifs [56]	65.2	67.1	-	-
GC	Split-1	MPS [19]	65.1	66.9	-	-
GC	Split-2	VTransE [67]	62.6	62.9	16.4	16.4
GC	Split-2	STA [53]	62.7	62.9	18.9	18.9
GC	Split-2	**ST-GAN-f$^{(w)}_{ref}$**	63.1	63.7	**19.0**	**19.0**
GC	Split-2	**ST-GAN-f$^{(s)}_{ref}$**	**63.6**	**64.0**	-	-
No GC	Split-2	DRL [6]	-	-	-	-
No GC	Split-2	DSR [27]	69.06	74.37	-	-
No GC	Split-2	**ST-GAN-f$^{(w)}_{ref}$**	**77.97**	**84.55**	**36.2**	**49.5**

Table 3: Relationship recognition accuracy (%) comparisons with graph constraints (GC) and without graph constraints (No GC). * indicates our implementation using their codes.

recognized by VTransE [67] but correctly recognized by ours. We also show the *cheap* object instances and relation reference images. In (a), our method failed to predicate the correct label *ear-on-teddy bear*. When observing the object instances, we see the reason is that most *ear* examples are *on* the heads of human or animals instead of toys.

5 Conclusions

In this work, we propose a novel relation feature generation model ST-GAN, in order to tackle recognition problems for rare and unseen relation categories. Our ST-GAN utilizes not only the real/fake discriminator but also the relationship classier, encouraging the generated feature to be useful for the relationship classification. Our approach transforms the *cheap* object instances to augment data for *expensive* compositional relations. Extensive experiments demonstrate our augmented data can be used to achieve superior performances, particularly for few-shot and zero-shot classes.

Acknowledgments

This research is part of NExT++ research which is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative, and it is part of CREATE program, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG. It is also partially supported by German Research Foundation (DFG CRC 1223), and by the Nvidia Corporation through the Memorandum of Understanding with the Advanced Robotics Centre of the National University of Singapore on autonomous systems technologies.

References

[1] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations (ICLR)*, 2018.

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[3] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2017.

[4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2017.

[7] Li Deng and Yang Liu. *Deep Learning in Natural Language Processing*. Springer, 2018.

[8] Ming Ding, Jie Tang, and Jie Zhang. Semi-supervised learning on graphs with generative adversarial nets. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, 2018.

[9] Mandar Dixit, Roland Kwitt, Marc Niethammer, and Nuno Vasconcelos. Aga: Attribute-guided augmentation. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2017.

[10] Rafael Felix, BG Vijay Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[12] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2015.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[14] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *International Conference on Learning Representations (ICLR)*, 2018.

[15] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[17] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Roei Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, and Amir Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[20] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[21] 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. 2016. URL https://arxiv.org/abs/1602.07332.

[22] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning (ICML)*, 2018.

[23] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[25] Yikang Li, Wanli Ouyang, Zhou Bolei, Shi Jianping, Zhang Chao, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[26] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[27] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *AAAI*, 2018.

[28] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[30] Yaoyao Liu, Qianru Sun, Anan Liu, Yuting Su, Bernt Schiele, and Tat-Seng Chua. Lcc: Learning to customize and combine neural networks for few-shot learning. *arXiv*, 1904.08479, 2019.

[31] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[33] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Liqian Ma, Qianru Sun, Bernt Schiele, and Luc Van Gool. A novel bilevel paradigm for image-to-image translation. *arXiv*, 1904.09028, 2019.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, 2013.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[37] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[38] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[39] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning convex combination of semantic embeddings. *International Conference on Learning Representations ICLR*, 2014.

[40] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *International Conference on Learning Representations (ICLR)*, 2017.

[41] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[42] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[43] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. *arXiv:1811.10696*, 2018.

[44] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *In International Conference on Learning Representations (ICLR)*, 2017.

[45] Oren Rippel, Manohar Paluri, Piotr Dollar, and Lubomir Bourdev. Metric learning with adaptive density discrimination. In *International Conference on Learning Representations (ICLR)*, 2016.

[46] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *Advances in neural information processing systems (NIPS)*, 2013.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[48] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[49] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems (NIPS)*. 2018.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[51] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[52] Jie Song, Chengchao Shen, Jie Lei, An-Xiang Zeng, Kairi Ou, Dacheng Tao, and Mingli Song. Selective zero-shot classification with augmented attributes. In *European Conference on Computer Vision (ECCV)*, 2018.

[53] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017*, 2017.

[54] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[55] Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*. Springer, 1998.

[56] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[57] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems (NIPS)*, 2006.

[58] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[59] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[60] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[61] Hsuan-Kung Yang, An-Chieh Cheng, Kuan-Wei Ho, Tsu-Jui Fu, and Chun-Yi Lee. Visual relationship prediction via label clustering and incorporation of depth information. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[62] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[63] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[64] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[65] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[66] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[67] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[68] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[69] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed M Elgammal. Relationship proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[70] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision (CVPR)*, 2015.

[71] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.