Resolving Semantic Confusions for Improved Zero-Shot Detection

Sandipan Sarma
sandipan.sarma@iitg.ac.in
Sushil Kumar
sushil13129340@gmail.com
Arijit Sur
arijit@iitg.ac.in

Indian Institute of Technology Guwahati
Guwahati, India

Abstract

Zero-shot detection (ZSD) is a challenging task where we aim to recognize and localize objects simultaneously, even when our model has not been trained with visual samples of a few target ("unseen") classes. Recently, methods employing generative models like GANs have shown some of the best results, where unseen-class samples are generated based on their semantics by a GAN trained on seen-class data, enabling vanilla object detectors to recognize unseen objects. However, the problem of semantic confusion still remains, where the model is sometimes unable to distinguish between semantically-similar classes. In this work, we propose to train a generative model incorporating a triplet loss that acknowledges the degree of dissimilarity between classes and reflects them in the generated samples. Moreover, a cyclic-consistency loss is also enforced to ensure that generated visual samples of a class highly correspond to their own semantics. Extensive experiments on two benchmark ZSD datasets – MSCOCO and PASCAL-VOC – demonstrate significant gains over the current ZSD methods, reducing semantic confusion and improving detection for the unseen classes. Codes and models will be released at https://github.com/sandipan211/ZSD-SC-Resolver.

1 Introduction

The idea of *zero-shot detection* (ZSD) [4, 8, 35, 61] has been recently introduced with the aim of transferring knowledge about some "seen" classes to the "unseen" with the help of semantic information relating these classes. At test time, trained ZSD models are evaluated in two settings – (i) test images contain unseen objects only (conventional ZSD), and (ii) test images can have objects from both seen and unseen classes (*generalized ZSD or GZSD*). While many of the ZSD approaches are inspired by the success of zero-shot recognition methods focusing on improving visual-semantic alignment [4, 8, 27, 34, 35], others tried leveraging additional information in the form of textual descriptions [23] and synthesizing unseen samples using generative methods [16, 58, 62], achieving state-of-the-art results. Interestingly, most of these approaches suffer from the problem of *semantic confusion*, where the knowledge transfer between the seen and unseen classes bridged by semantic representations is not discriminative enough at times to distinguish between semantically-similar classes. This results in low average precisions for the ZSD models, which could prove to be catastrophic when deployed in real-world environments in the future, *e.g.* in medical imaging systems where a high false-positive rate can negatively impact their reliability, or in underwater explorations where low luminance and turbidity can misguide the detection of marine debris.

We propose a generative method for ZSD, which aims to resolve semantic confusion by utilizing a triplet loss [44] while training the feature synthesizer. Specifically, we use Faster-RCNN [37] as a backbone object detector that can be trained on images containing only seen class objects. Fixed-size feature vectors for these objects are used to train a conditional Wasserstein GAN [3] (cWGAN) regularized by a classification loss, following the success of [53] in the zero-shot recognition task. In order to ensure diversity among the synthesized features, we use a regularization term [26] that alleviates the issue of mode collapse in conditional GANs. However, such a cWGAN only learns how to synthesize image features conditioned upon class semantics and does not account for the degree of dissimilarity between object classes while learning to synthesize their features. Hence we introduce a triplet loss that can primarily help in learning discriminative features for the semantically-similar classes, resolving semantic confusion whenever these synthesized features are utilized for our detection pipeline ahead. Moreover, we explicitly aim to maintain the consistency between the synthesized visual features and semantics of the corresponding class by incorporating a cyclic-consistency loss enforcing the synthesized visual features to reconstruct their semantics. The trained cWGAN is used to generate unseen class features, which are used to update the classifier of the pretrained Faster-RCNN, empowering it to detect unseen-class objects as well. Since the performance of this classifier is directly related to the quality of synthesized features used as inputs for training, accounting for inter-class dissimilarity and visual-semantic consistency can impact the performance of the detector. Moreover, using a generative method also minimizes the *hubness problem* [9, 15, 16, 33, 45, 58].

We summarize our contributions in this work as follows: (i) we propose using a triplet loss with a flexible semantic margin (refer to Sec. 3.2) while training a feature synthesizer (cWGAN) which generates unseen object features conditioned upon class semantics and enables our backbone object detector to detect both seen and unseen objects, resolving semantic confusion between similar objects; (ii) visual-semantic consistency is maintained during feature generation, ensuring generated features correspond well to their semantic counterparts; (iii) extensive experiments are performed on two benchmark ZSD datasets (MSCOCO and PASCAL-VOC) which show that our method remains comparable to the best existing methods in case of PASCAL-VOC, but comprehensively beats these methods on the more challenging and bigger dataset MSCOCO in both conventional ZSD and GZSD settings.

2 Related Work

Motivated by zero-shot classification (ZSC) methods [1, 2, 6, 11, 12, 18, 20, 22, 25, 29, 30, 31, 38, 39, 40, 43, 46, 47, 49, 51, 53, 54, 57, 63], the more challenging task of ZSD [4, 8, 35, 61] started gaining attention since 2018. The initial works seek to improve visual-semantic alignment and extend popular ZSC frameworks like ConSE [31] with trusted object detectors like Faster-RCNN. In these works, usually projection functions are learned [8, 27, 35] for capturing seen-unseen and visual-semantic relationships. However, recent methods have shown that such projection-based strategies can be improved. Contrastive losses designed with respect to semantic vectors and put into action within a joint intermediate embedding space for the visual features and semantic vectors in [56], and a polarity loss [34] which refines the noisy semantic vectors and explicitly maximizes the gap between positive and negative predictions, are two examples of such methods.

Some methods additionally target the problem of background-unseen confusion (BUC), where ZSD models confuse unseen objects with background at test time due to low objectness scores for unseen objects. Additional data from external sources is used for obtaining a vocabulary with classes belonging to neither seen nor unseen classes in [4], encoding an idea about the background classes. Vocabulary atoms [7] enrich the semantic space with a diverse set of linguistic concepts and help relate to the visual features better in [34]. The same vocabulary is used in [60] along with a background-learnable RPN for detecting objects.

Rather than focusing only on visual-semantic alignment, several methods explore some other limitations in ZSD using different data structures and multi-modal approaches. Unseen-class *localization* gets priority in [63], where predicting class attributes is a side-task and the produced bounding boxes utilize both visual and semantic information. In a multi-modal approach, [23] uses unit-level and word-level attention from a language branch for weighing the outputs of a visual branch for detecting objects. A GCN-based [50] approach is taken up in [55], utilizing a graph construction module and two semantics-preserving graph propagation modules. Transformer-based encoder-decoder networks have been used in [59], achieving a stronger ability to deal with BUC and recall unseen objects.

With generative networks [13, 21] minimizing the hubness problem [33], a few ZSD methods have also turned to such networks. A conditional VAE is employed in [62] for synthesizing unseen features used to update the confidence predictor of a YOLO [36] detector, pre-trained with seen objects. In another work [58], three separate GANs are used to generate visual features with both intra-class variance and IoU variance. Instead, [16] encourages a unified GAN model that generates discriminative object features and ensures feature diversity. Inspired by the potential of generative methods, we also employ one for our ZSD model (Fig. 1). However, our work differs majorly from the previous generative approaches in the use of visual-semantic cyclic consistency loss and in addressing the problem of semantic confusion, which none of the previous works do. We compare our results in Sec. 4 with methods following all these aforementioned approaches.

3 Method

Problem Setting: We formally define ZSD here. Let $\mathcal{C}^s = \{1, 2, \ldots S\}$ and $\mathcal{C}^u = \{S+1, S+2, \ldots S+U\}$ be label sets of S number of seen and U number of unseen classes respectively, such that $\mathcal{C}^s \cap \mathcal{C}^u = \phi$. Moreover, in object detection, a concept of *background class* must be identified too – so total class labels become $S+U+1$. Let the training data $\mathcal{D}^{tr} = \{I_m, \{O_m^i\}_{i=1}^{N_m}\}_{m=1}^{M}$ consist of M images, having N_m objects with class annotations in the set $\{O_m^i\}_{i=1}^{N_m}$ for an image I_m. The i^{th} object in I_m is annotated as $O_m^i = \{B_m^i, c_m^i\}$, where $B_m^i = \{x_m^i, y_m^i, w_m^i, h_m^i\}$ denotes the bounding-box coordinates and $c_m^i \in \mathcal{C}^s$. Moreover, the semantic descriptions for the seen and unseen classes (d-dimensional word embeddings [19, 28, 32]) are given as $\mathcal{P}^s \in \mathbb{R}^{S \times d}$ and $\mathcal{P}^u \in \mathbb{R}^{U \times d}$ respectively. At test time, images containing objects from both $\mathcal{C}^s$ and $\mathcal{C}^u$ can be given, and the goal would be to predict bounding boxes for every foreground object, along with their class labels.

Backbone object detector: We use a Faster-RCNN [37] $\Phi_{\texttt{frcn}}$ with a ResNet-101 [17] pretrained on ImageNet [41] classes (except the *overlapping unseen classes* [52]) as a feature extractor for the input images, yielding a convolutional feature map (*ConvMap*). A *Region*

Proposal Network (RPN) predicts objectness scores for different *region proposals* in the ConvMap. For each of the top N_p proposals adjudged as *foreground* by the RPN, fixed-size feature vectors are extracted from its projection on the ConvMap via RoI pooling. These are fed to fully-connected layers that branch into a classification module $\phi_{\texttt{frcn-c}}$ which classifies a proposal into one of $N+1$ classes (N foreground classes and a "background" class), and a bounding-box regression module $\phi_{\texttt{frcn-r}}$ which regresses over box coordinates for object localization. $\Phi_{\texttt{frcn}}$ is trained on $\mathcal{D}^{tr}$, and used to extract object-level visual features $f_m^i \in \mathcal{F}_m$ for every seen object in an image I_m.

3.1 The feature synthesizer

The baseline architecture of the proposed model is depicted in Fig. 1 (similar GAN also used in a prior ZSC task [53]). The collection of object features from all training images $\mathcal{F}^s = \{\mathcal{F}_m\}_{m=1}^M$ along with the corresponding seen-class labels serve as *real data* for a conditional WGAN [3] (cWGAN). A generator network learns a mapping function $\mathcal{G} : \mathcal{Z} \times \mathcal{P}^s \longrightarrow \mathcal{F}^s$ that takes $p \in \mathcal{P}^s$ and $z \in \mathcal{Z}$ as inputs and learns the underlying distributions of the visual features from $\mathcal{F}^s$, relates them to the corresponding semantics. Here, $z \sim \mathcal{N}(0,1) \in \mathbb{R}^{z_d}$ is a random noise vector sampled from a Gaussian distribution. During training, the generator $\mathcal{G}$ generates class-wise seen object features (*fake data*), feeds them to a critic network $\mathcal{Q}$ for scoring its *realness* or *fakeness* and recalibrates itself based on the feedback from $\mathcal{Q}$ to make the generated feature distribution as close to real distribution as possible, minimizing the loss:

$$\mathcal{L}_{\texttt{WGAN}} = \mathbb{E}[\mathcal{Q}(f,c)] - \mathbb{E}[\mathcal{Q}(\tilde{f},c)] + \lambda \mathbb{E}[(||\nabla_{\hat{f}} \mathcal{Q}(\hat{f},c)||_2 - 1)^2] \tag{1}$$

where the first two terms represent the critic loss in WGAN and the third term represents a gradient penalty [14], with λ being the penalty coefficient. $\tilde{f} = \mathcal{G}(z,p)$ denotes the generated feature, and $\hat{f} = \rho f + (1-\rho)\tilde{f}$, with $\rho \sim U(0,1)$ [14, 53]. We add a regularization term [53] that enforces discriminative feature generation using:

$$\mathcal{L}_{\texttt{CLS}} = -\mathbb{E}[\log \mathbf{p}(c|\mathcal{G}(z,p); \phi_{cls}^{sm})] \tag{2}$$

where $\mathbf{p}(y|\mathcal{G}(.))$ denotes a classification probability given by a linear softmax classifier ϕ_{cls}^{sm} pretrained on $\mathcal{F}^s$. Additionally, inspired by [16], we consider the impact of individual noise vectors on feature generation and enhance feature diversity to prevent the problem of mode collapse [42] by including a mode-seeking regularization term [26]:

$$\mathcal{L}_{\texttt{MS}} = \mathbb{E}[||\mathcal{G}(z_1,p) - \mathcal{G}(z_2,p)||_1 / ||z_1 - z_2||_1] \tag{3}$$

3.2 Triplet loss for resolving semantic confusion

Semantic confusion is a major hindrance in ZSL tasks, where semantically-similar classes are hard to distinguish at test time, dropping prediction scores for the seen and unseen classes. In the ZSC task, this problem has been addressed using triplet loss-based methods [1, 12]. In this work, we leverage a modified triplet loss for our ZSD task. However, our execution of the triplet loss is different from the ones used in ZSC, where multi-modal

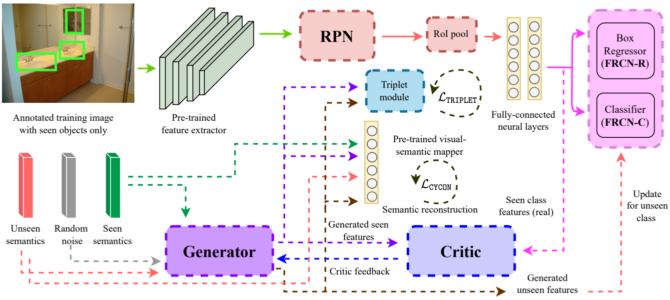

Figure 1: **Model architecture for the proposed ZSD model.** The *solid* arrows show the workflow in the backbone object detector part, whereas the *dashed arrows* show the workflow regarding the feature synthesizer (cWGAN), optimizing the objective function in Eq. 6.

triplets are formed with visual features as "anchors" and class semantics as "positive" or "negative" matches, and compatibility scores indicate multi-modal similarity. On the contrary, our triplets are strictly visual feature-based, having the form $<\tilde{f}_a, \tilde{f}_p, \tilde{f}_n>$, where $\tilde{f}_a$ is the anchor feature of a cWGAN-generated sample from a given class, $\tilde{f}_p$ is a positive match of the same class as $\tilde{f}_a$, and $\tilde{f}_n$ is a negative match of a different class. Instead of compatibility scores, we focus on similarity in the visual space spanned by our triplets and optimize:

$$\mathcal{L}_{\texttt{TRIPLET}} = \max\{0, (d(\tilde{f}_a, \tilde{f}_p) - d(\tilde{f}_a, \tilde{f}_n) + \Delta)\} \tag{4}$$

where $d(.,.)$ is the Euclidean distance. Our motivation is that the quality of the features generated on the basis of class semantics heavily impacts the classification ability of $\phi_{\texttt{frcn-c}}$ at a later stage – therefore, accounting for inter-class dissimilarities explicitly in the visual space can address semantic confusion during the feature generation phase itself. In addition, Eq. 4 uses a flexible semantic margin Δ acquired as a pre-computed value from a rescaled Mahalanobis distance matrix with respect to class semantics [5], instead of keeping it at a fixed, constant value [1, 12]. This is because we want the generator to acknowledge the degree of dissimilarity between two classes and reflect them in the generated features by accounting for the first and second-order statistics of the class semantics.

To the best of our knowledge, such an approach has not been considered in ZSD yet. Moreover, we recognize the fact that utilizing triplet loss for a generative ZSD approach like ours could prove quite beneficial as our cWGAN can explicitly learn to reduce confusion between semantically-similar classes, producing generated features that are robust to such semantic similarities.

3.3 Cyclic consistency between the visual and semantic spaces

The cWGAN assumes that generated object features for a class conditioned upon its semantics would have a distribution close to the real features of that class. This stems from an implicit assumption that the visual and semantic distributions for that class are relatively similar. However, this can bias the generated features of the unseen classes towards semantically-similar seen classes on which the cWGAN is trained, negatively affecting $\phi_{\texttt{frcn-c}}$ at the next

Table 1: ZSD and GZSD performance of various methods on MSCOCO in terms of mAP and Recall@100 (RE@100) at an IoU threshold of 0.5. HM denotes the harmonic mean of seen and unseen results for GZSD. The best results and second-best results are shown in *red* and *blue* respectively. Results are shown for the 65/15 split of MSCOCO.

Metric	Method	ZSD	GZSD		
			Seen	Unseen	HM
mAP	PL [34]	12.40	34.07	12.40	18.18
	BLC [60]	14.70	36.00	13.10	19.20
	ACS-ZSD [27]	15.34	-	-	-
	SUZOD[16]	17.30	37.40	17.30	23.65
	ZSDTR [59]	13.20	**40.55**	13.22	20.16
	ContrastZSD [56]	18.60	40.20	16.50	23.40
	Ours	**20.10**	37.40	**20.10**	**26.15**
RE@100	PL [34]	37.72	36.38	37.16	36.76
	BLC [60]	54.68	56.39	51.65	53.92
	ACS-ZSD[27]	47.83	-	-	-
	SUZOD[16]	61.40	58.60	60.80	59.67
	ZSDTR [59]	60.30	**69.12**	59.45	61.12
	ContrastZSD [56]	59.50	62.90	58.60	60.70
	Ours	**65.10**	58.60	**64.00**	**61.18**

stage. Hence, we implement a cyclic consistency module forcing the generated features to reconstruct the class semantics based on which they were generated in the first place. A visual-semantic mapper $\mathcal{M} : \mathcal{F}^s \longrightarrow \mathcal{P}^s$ is first trained on seen data, which is an MLP that learns to map object features of seen classes to their semantic counterparts, minimizing a semantic reconstruction loss. During cWGAN training, the synthesized features are passed to the pretrained $\mathcal{M}$ which reconstructs the semantics from these features, and a reconstruction loss is optimized as:

$$\mathcal{L}_{\texttt{CYCON}} = \mathop{\mathbb{E}}_{p\sim\mathcal{P}^s}[||p-\mathcal{M}(\mathcal{G}(z,p))||_2^2] + \mathop{\mathbb{E}}_{p\sim\mathcal{P}^u}[||p-\mathcal{M}(\mathcal{G}(z,p))||_2^2] \tag{5}$$

The overall objective function becomes:

$$\min_{\mathcal{G}} \max_{\mathcal{Q}} \alpha_1 \mathcal{L}_{\texttt{WGAN}} + \alpha_2 \mathcal{L}_{\texttt{CLS}} + \alpha_3 \mathcal{L}_{\texttt{MS}} + \alpha_4 \mathcal{L}_{\texttt{CYCON}} + \alpha_5 \mathcal{L}_{\texttt{TRIPLET}}, \tag{6}$$

making the generated features diverse and robust to semantic confusions and visual-semantic inconsistencies. Once the cWGAN is trained, we train a classifier ϕ_{cls}^u using the generated unseen features. The learned weights are provided to $\phi_{\texttt{frcn-c}}$ to make it capable of classifying unseen visual features.

4 Experiments

4.1 Datasets and evaluation metrics

We evaluate our proposed method extensively on two commonly used datasets in ZSD – MSCOCO and PASCAL-VOC. MSCOCO [24] is a large-scale dataset containing annotated

Table 2: Class-wise average precisions (APs) on unseen classes (ZSD) from MSCOCO with an IoU threshold of 0.5. The best results and second-best results are shown in *red* and *blue*.

Method	**Overall**	aeroplane	train	p.meter	cat	bear	suitcase	frisbee	snowbrd.	fork	sandwich	hot dog	toilet	mouse	toaster	hair drier
PL[34]	12.40	20.0	48.2	0.63	28.3	13.8	12.4	**21.8**	15.1	8.9	8.5	0.87	5.7	0.04	**1.7**	0.03
ACS-ZSD[27]	15.34	8.72	25.5	**6.59**	40.8	54.0	9.55	10.6	26.8	16.4	11.0	4.99	7.83	**6.21**	1.32	0.0
SUZOD[16]	17.30	17.8	46.3	0.7	63.1	41.0	10.5	0.7	30.2	**16.5**	17.6	0.0	13.4	1.6	0.4	**0.2**
Ours	**20.10**	**22.9**	**53.3**	0.6	**64.9**	**54.3**	**13.2**	1.2	**31.2**	15.7	**22.6**	0.0	**17.5**	2.7	0.7	**0.2**

Table 3: mAP (in %) for PASCAL-VOC dataset. The unseen classes are shown in italic. The best and second-best results are shown in *red* and *blue*.

Method	**seen**	**unseen**	aeroplane	bicycle	bird	boat	bottle	bus	cat	chair	cow	d.table	horse	motorbike	person	p.plant	sheep	tvmonitor	*car*	*dog*	*sofa*	*train*
HRE [8]	65.6	54.2	70.0	73.0	76.0	54.0	42.0	**86.0**	64.0	40.0	54.0	**75.0**	80.0	80.0	75.0	34.0	69.0	**79.0**	55.0	82.0	55.0	26.0
PL [34]	63.5	62.1	74.4	71.2	67.0	50.1	50.8	67.6	84.7	44.8	68.6	39.6	74.9	76.0	79.5	39.6	61.6	66.1	**63.7**	87.2	53.2	44.1
SAN [33]	69.6	57.6	71.4	78.5	74.9	61.4	48.2	76.0	89.1	51.1	**78.4**	61.6	84.2	76.8	76.9	42.5	71.0	71.7	56.2	85.3	**62.6**	26.4
BLC [56]	**75.1**	55.2	78.5	83.2	77.6	**67.7**	**70.1**	75.6	87.4	55.9	77.5	71.2	**85.2**	82.8	77.6	**56.1**	**77.1**	78.5	43.7	86.0	60.8	30.1
SUZOD [16]	74.7	**63.1**	**80.6**	**84.2**	**78.7**	66.7	67.6	74.0	**91.5**	**61.1**	75.0	63.4	82.6	**85.7**	**84.9**	47.8	76.9	74.8	55.2	92.3	59.0	**46.1**
Ours	74.7	62.7	80.4	84.1	**78.7**	66.6	67.6	73.8	91.4	**61.1**	75.0	63.7	83.1	85.5	**84.9**	47.3	76.7	74.7	55.6	**92.6**	57.5	45.0

images from 80 object classes, for which we use the seen/unseen split of 65/15 provided by [34]. PASCAL-VOC 2007/2012 [10] contains annotated images from 20 object classes, for which we use the 16/4 split provided in [8]. We follow [16] while obtaining the sets of training and testing images. Following [4] and [34], we use Recall@100 (RE@100) and mean average precision (mAP) as evaluation metrics and report our results at an IoU of 0.5. In the GZSD setting, we follow [34] and show the mAP on seen and unseen classes and consider their harmonic mean (HM) as the overall performance metric.

4.2 Implementation details

For the RPN, anchor bounding boxes with an IoU ≥ 0.7 are regarded as *foreground* , whereas those with IoU ≤ 0.3 are regarded as *background*, yielding $N_p = 2000$ proposals for each image with an NMS threshold of 0.7. The backbone Faster-RCNN model is trained on seen data first for 12 epochs and 4 epochs on the GPU for MSCOCO and PASCAL respectively. Classifiers and bounding-box regressors are both fully-connected neural layers. The cWGAN takes 300-dimensional FastText embedding vectors [19] as class semantics and 1024-dimensional object features extracted using RoI pooling layer of the Faster-RCNN (pretrained on seen data) as *real features*. The generator and critic networks are implemented as single-layered neural networks with 4096 hidden units. For Eq. 6, we empirically set the hyperparameter values as $\alpha_1 = 1.0, \alpha_2 = 0.01, \alpha_3 = 0.01, \alpha_4 = 0.01$ and $\alpha_5 = 0.1$. The cWGAN is trained for 55 epochs, and the weights from the best epoch are used further. Adam optimizer is used for both $\mathcal{G}$ and $\mathcal{Q}$ networks, with a learning rate of 0.0005 and mini-batch size of 128. Generated features for both seen and unseen classes are checked for their consistency via Eq. 5, and also used for constructing all possible triplets *online* [44] for Eq. 4.

4.3 Results: Comparison with the State-of-the-art

Results on MSCOCO. (i) *ZSD setting*: Table 1 shows that our method achieves a **relative mAP gain of 8%** over the next-best method [56], and a **relative RE@100 gain of 6%** over the next-best method [16]. This reflects the superior performance of generative methods over others, especially because we get rid of the hubness problem while augmenting visual data for the unseen classes in a cycle-consistent manner.

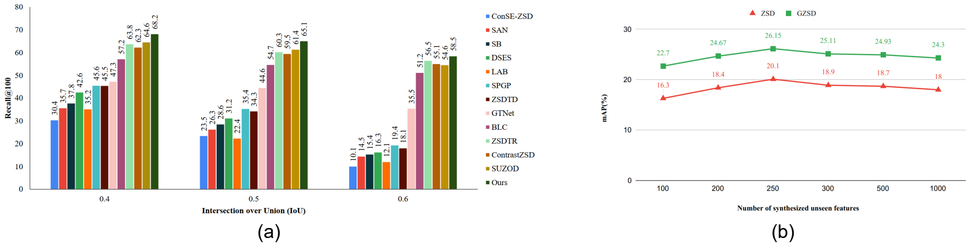

Figure 2: (a) Comparison of recall@100 for various IoU thresholds on MSCOCO; (b) mAP variation depending on the number of synthesized unseen features in case of MSCOCO (unseen mAP considered for ZSD and HM of seen and unseen mAP considered for GZSD).

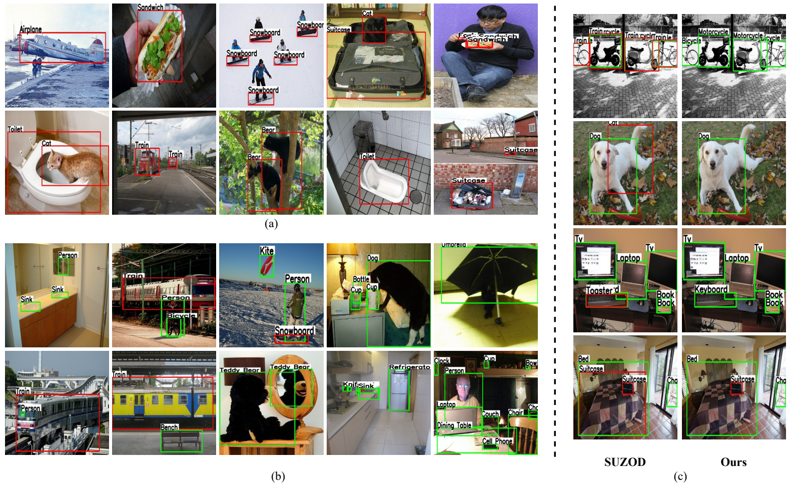

Figure 3: Qualitative results on MSCOCO (best viewed in zoom). (a) ZSD results; (b) GZSD results; (c) Comparison of our GZSD results with the current state-of-the-art (SUZOD [16]). The seen and unseen class objects are detected within *green* and *red* boxes.

Table 4: Impact of including different loss terms for training the feature synthesizer on mAP (in %) during evaluation on the MSCOCO dataset.

$\mathcal{L}_{\texttt{WGAN}}$	$\mathcal{L}_{\texttt{MS}}$	$\mathcal{L}_{\texttt{CLS}}$	$\mathcal{L}_{\texttt{CYCON}}$	$\mathcal{L}_{\texttt{TRIPLET}}$	mAP
✓	✓	✓	✓		18.4
✓		✓	✓	✓	18.5
✓	✓		✓	✓	19.3
✓	✓	✓		✓	19.4
✓	✓	✓	✓	✓	**20.1**

Moreover, the gain we achieve for a challenging metric like mAP, where false positives are penalized, suggests that our triplet loss is indeed reducing semantic confusion and helping in correct classification of unseen objects. At a class-level (Tab. 2), we achieve better results for most unseen classes. However, some classes like *parking meter*, *frisbee* and *hot dog* show decreased performance, possibly due to the unavailability of semantically-similar seen classes, affecting conditional feature generation. The same is reflected in the t-SNE [48] plot in Fig. 5(a), where these classes do not form very well-defined clusters as compared to other unseen classes like *bear*, pointing out a challenging aspect of generative approaches.

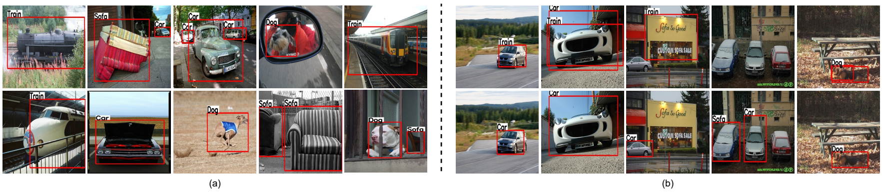

Figure 4: Qualitative results on PASCAL-VOC (best viewed in zoom). (a) Unseen detections; (b) Comparison of our results (bottom row) with SUZOD [16] (top row).

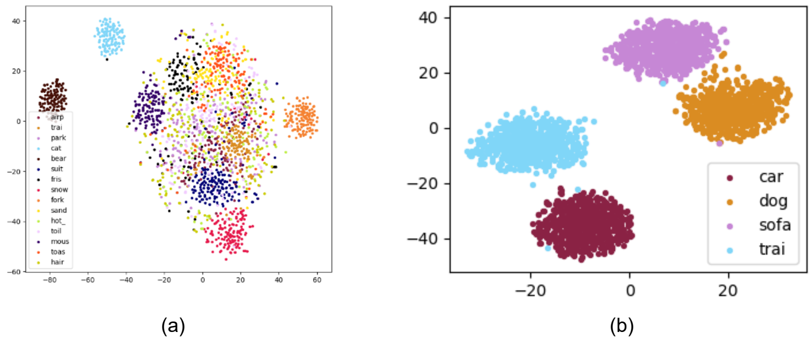

Figure 5: t-SNE visualization [48] of the generated features for unseen classes in (a) MSCOCO and (b) PASCAL-VOC.

Figure 2(a) shows a comparison with different methods for the RE@100 values at different IoU thresholds during ZSD evaluation. Our method outperforms all methods [4, 16, 23, 31, 35, 55, 56, 58, 59, 60] for three IoU thresholds, even when some of the existing methods use additional semantic information from external vocabularies [4, 23].

(ii) *GZSD setting*: GZSD is a challenging setting as there always exists a bias towards the seen classes for the model. Table 1 shows a **relative mAP gain of 10.5%** over the next-best method (SUZOD [16]) considering HM. We, therefore, comprehensively beat the current state-of-the-art results for MSCOCO in both ZSD and GZSD settings.

(iii) *Qualitative analysis*: Figure 3(a) shows our method can detect multiple unseen instances of the same object class as well as multiple objects from different classes. Objects from different viewpoints and small sizes (Fig. 3(b)) have also been detected well. Low localization error is encountered for GZSD cases too, with good robustness against background clutter and occlusion. Figure 3(c) compares the localization and classification abilities of our method with the state-of-the-art GZSD method [16]. Unlike our method, [16] succumbs to semantic confusion and wrongly detects objects like *bicycle, motorcycle, keyboard*, and *bed*.

Results on PASCAL-VOC. The unseen classes show improved results as compared to state-of-the-art (SUZOD [16]) on account of improvement in the generated unseen-class features (Fig. 4(a)). However, Tab. 3 reports our unseen mAP is second-best to [16] (only by 0.4%), probably due to the relatively small size of the training data, which might not be enough to learn inter-class dissimilarities sufficiently and hence make an impact on the feature synthesizer and object detector. Nevertheless, we compare our visual results with [16] in Fig. 4, showing that semantic confusion between *car* and *train* is common for [16], but not for us. The t-SNE plot of the generated unseen class features in Fig. 5(b) demonstrates good separability between classes, even for semantically-similar classes like *car* and *train*.

4.4 Ablation Studies

Effect of the loss components: Table 4 shows that $\mathcal{L}_{\texttt{CYCON}}$ provides an mAP boost as per our intuition, which indicates that generated features by the cWGAN are more discriminative and consistent with their class semantics. However, $\mathcal{L}_{\texttt{TRIPLET}}$ has the strongest influence on ZSD, without which mAP is 18.4% – but jumps to 19.4% even when $\mathcal{L}_{\texttt{CYCON}}$ is not utilized. $\mathcal{L}_{\texttt{CLS}}$ provides mAP boost only when included in conjunction with the other loss terms. However, optimal performance is attained when using all five loss terms for training cWGAN.
Effect of the number of generated examples: We fix the number of generated seen features while training cWGAN and vary the number of class-wise unseen features generated by the trained cWGAN. When evaluated on MSCOCO in ZSD and GZSD settings, we find our model achieves optimal results when 250 features are generated per unseen class (Fig. 2(b)).

5 Conclusion

While transferring knowledge from the seen to unseen classes, most existing ZSD methods face confusion while detecting semantically similar classes. We propose a generative method that inherently eliminates the hubness problem in zero-shot conditions. Our triplet loss with a flexible semantic margin acknowledges the degree of dissimilarity between object classes while learning to synthesize discriminative object features. Moreover, a cyclic-consistency loss is enforced to maintain the visual-semantic consistencies during feature generation. Experiments and ablation studies on two challenging datasets show that we achieve state-of-the-art results, both qualitatively and quantitatively, improving upon some of the fundamental challenges the existing ZSD methods face, such as semantic confusion, high false-positive rate, and misclassification of localized objects. From a research perspective, future directions include improving the model architecture for better localization of unseen classes and reducing background-unseen confusion. From an application perspective, our ZSD model can be used as a plug-and-play module in the future for various vision applications, even in challenging environments. For instance, images captured in underwater environments can be pre-processed via image restoration techniques and fed to our ZSD model. Our model can detect novel species of fish, corals, and also help in trash detection by localizing different kinds of unseen trash objects such as plastic snack wrappers.

6 Acknowledgement

This work is supported by IITG Technology Innovation and Development Foundation (IIT-GTI&DF), which has been set up at IIT Guwahati as a part of the National Mission on Interdisciplinary Cyber Physical Systems (NMICPS). IITGTI&DF is undertaking research, development and training activities on Technologies for Under Water Exploration with the financial assistance from Department of Science and Technology, India through grant number DST/NMICPS/TIH12/IITG/2020. Authors gratefully acknowledge the support provided for the present work. We also acknowledge the Department of Biotechnology, Govt. of India for the financial support for the project BT/COE/34/SP28408/2018 (for computing resources).

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936, 2015.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[4] 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)*, pages 384–400, 2018.

[5] Yannick Le Cacheux, Herve Le Borgne, and Michel Crucianu. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10333–10342, 2019.

[6] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5327–5336, 2016.

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[8] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *BMVC*, 2018.

[9] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[11] Liangjun Feng and Chunhui Zhao. Transfer increment for generalized zero-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2506–2520, 2021.

[12] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[15] Dikshant Gupta, Aditya Anantharaman, Nehal Mamgain, Vineeth N Balasubramanian, CV Jawahar, et al. A multi-space approach to zero-shot object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1209–1217, 2020.

[16] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[17] 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*, pages 770–778, 2016.

[18] He Huang, Yuanwei Chen, Wei Tang, Wenhao Zheng, Qing-Guo Chen, Yao Hu, and Philip Yu. Multi-label zero-shot classification by learning to transfer from external knowledge. *arXiv preprint arXiv:2007.15610*, 2020.

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2068.

[20] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, and Osamu Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*, pages 3657–3664, 2012.

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

[22] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3174–3183, 2017.

[23] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8690–8697, 2019.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[25] Yu Liu and Tinne Tuytelaars. A novel baseline for zero-shot learning via adversarial visual-semantic embedding. *Proceedings BMVC 2020*, 2020.

[26] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1429–1437, 2019.

[27] Qiaomei Mao, Chong Wang, Shenghao Yu, Ye Zheng, and Yuqi Li. Zero-shot object detection with attributes-based category similarity. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(5):921–925, 2020.

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[29] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *CVPRW*, pages 2188–2196, 2018.

[30] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *ECCV*, 2020.

[31] Mohammad Norouzi, Tomás Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010.

[34] Shafin Rahman, Salman Khan, and Nick Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018.

[35] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, pages 547–563. Springer, 2018.

[36] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[38] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, pages 910–917, 2010.

[39] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, pages 1641–1648, 2011.

[40] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR, 2015.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[43] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[45] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 135–151. Springer, 2015.

[46] Chenwei Tang, Zhenan He, Yunxia Li, and Jiancheng Lv. Zero-shot learning via structure-aligned generative adversarial network. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2021.

[47] William Thong and Cees G.M. Snoek. Bias-awareness for zero-shot learning the seen and unseen. In *BMVC*, 2020. URL https://arxiv.org/abs/2008.11185.

[48] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[49] Maunil R Vyas, Hemanth Venkateswara, and Sethuraman Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *European Conference on Computer Vision*, pages 70–86. Springer, 2020.

[50] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

[51] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 69–77, 2016.

[52] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[53] 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*, pages 5542–5551, 2018.

[54] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-VAEGAN-D2: A feature generating framework for any-shot learning. In *CVPR*, pages 10267–10276, 2019.

[55] Caixia Yan, Qinghua Zheng, Xiaojun Chang, Minnan Luo, Chung-Hsing Yeh, and Alexander G Hauptman. Semantics-preserving graph propagation for zero-shot object detection. *IEEE Transactions on Image Processing*, 29:8163–8176, 2020.

[56] Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-guided contrastive network for zero-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[57] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015.

[58] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang. Gtnet: Generative transfer network for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12967–12974, 2020.

[59] Ye Zheng and Li Cui. Zero-shot object detection with transformers. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 444–448. IEEE, 2021.

[60] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[61] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):998–1010, 2019.

[62] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11693–11702, 2020.

[63] Yizhe Zhu, Jianwen Xie, Bingchen Liu, and Ahmed Elgammal. Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9844–9854, 2019.