Robust Region Feature Synthesizer for Zero-Shot Object Detection

Peiliang Huang[1], Junwei Han[1]*, De Cheng[2], Dingwen Zhang[1]*
[1]Brain and Artificial Intelligence Lab, School of Automation, Northwestern Polytechnical University
[2]State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University
{peilianghuang2017, junweihan2010, zhangdingwen2006yyy}@gmail.com, dcheng@xidian.edu.cn

Abstract

Zero-shot object detection aims at incorporating class semantic vectors to realize the detection of (both seen and) unseen classes given an unconstrained test image. In this study, we reveal the core challenges in this research area: how to synthesize robust region features (for unseen objects) that are as intra-class diverse and inter-class separable as the real samples, so that strong unseen object detectors can be trained upon them. To address these challenges, we build a novel zero-shot object detection framework that contains an Intra-class Semantic Diverging component and an Inter-class Structure Preserving component. The former is used to realize the one-to-more mapping to obtain diverse visual features from each class semantic vector, preventing miss-classifying the real unseen objects as image backgrounds. While the latter is used to avoid the synthesized features too scattered to mix up the inter-class and foreground-background relationship. To demonstrate the effectiveness of the proposed approach, comprehensive experiments on PASCAL VOC, COCO, and DIOR datasets are conducted. Notably, our approach achieves the new state-of-the-art performance on PASCAL VOC and COCO and it is the first study to carry out zero-shot object detection in remote sensing imagery.

1. Introduction

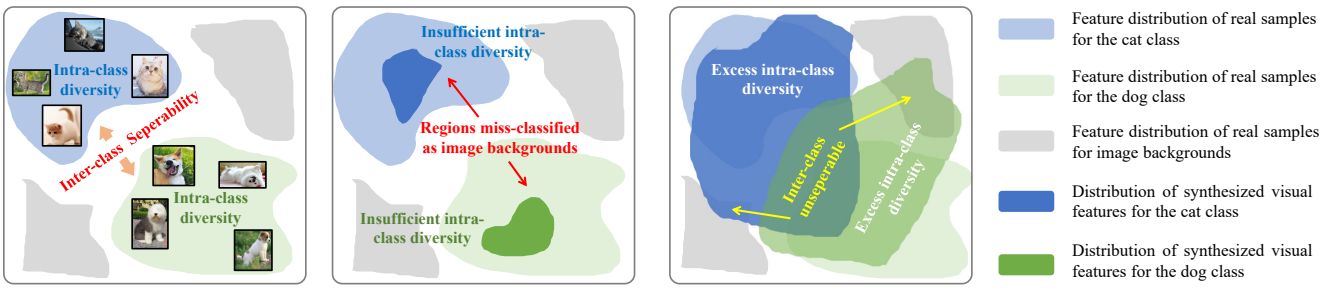
With the rapid development of the deep learning technologies, such as CNN [14, 34, 48] and Transformer [6, 26], great progresses have been made in the research field of object detection. Although the detection performance achieved by existing methods looks promising and encouraging, it exists a hidden drawback for applying them in real-world scenarios—The mainstream detection approaches have the strict constraint on the category to detect. Once the model is trained, it can only recognize objects that appear in the training data, whereas other objects appearing in the test images but unseen during training would confuse the model dramatically, leading to avoidless faults in detection results. To address this problem, the task of zero-shot object detection (ZSD) [4, 17, 33, 51] was raised in recent years. The goal is to enable the detection models to predict unseen objects which are without any available samples during training.

Earlier efforts on zero-shot object detection (ZSD) [4,33] focus on mapping function-based methods, which learn mapping functions from the visual space to the semantic space. With the learned mapping functions, unseen object categories can be predicted by mapping their visual features into the semantic space and then performing the nearest neighbor search in the semantic space. However, due to that the mapping functions are learned all upon the seen categories provided by the training data, the models would get significantly biased towards the seen categories when dealing with the visual features in testing [17]. Recently, generative model-based methods [17, 51] are presented as an alternative solution. Usually, these methods utilize generative models to synthesize visual features from the provided semantic embeddings [2, 30] corresponding to each object category. The synthesized visual features can then be used for training a standard detector for unseen classes. Generative model-based methods show stronger performance compared with mapping function-based methods in solving the bias problem as, although the samples corresponding to the unseen objects are still absent, the detectors are trained with synthesized visual features for the unseen objects.

However, the current generative model-based methods mainly follow the ideas presented in zero-short classification frameworks, such as [31, 41], where the synthesized visual features may perform well in the less complex classification scenarios but are not robust enough to obtain satisfying results in the complicated detection scenarios. To our best knowledge, there are two-fold challenges for synthesizing visual features for detection scenarios:

- **Intra-class diversity:** Objects in real-world detection scenarios present high variation in pose, shape, texture, etc., and one object instance may be covered by

*Corresponding author.



(a) Feature distribution of real samples

(b) Distribution of the synthesized visual features (Case I)

(c) Distribution of the synthesized visual features (Case II)

Figure 1. Illustrations of the problem studied in this work. In real cases, the feature space built by the samples show high intra-class diversity but still with inter-class separability like in (a), whereas the spaces of the synthesized visual features learned by existing approaches either have insufficient intra-class diversity, as shown in (b), or have excess intra-class diversity to make inter-class inseparable, as shown in (c).

several bounding boxes with different sizes and locations. This leads to the high diversity in their feature representations.

- **Inter-class separability:** Though having such variations, each object category still has easy-to-recognized characteristics that are distinct from other object categories as well as the image backgrounds, making the feature representations from different classes (including the background class) highly separable.

Although some existing approaches have recognized the importance of intra-class diversity [17, 49], without jointly considering the inter-class separability, these methods would either impose insufficient diversity to the synthesized visual features, leading to miss-classify the real unseen objects as image backgrounds (see Fig 1 (b)), or go too far to make the visual features synthesized for different class semantics mixed together, thus making the learned detection models obtain inaccurate object categories for foreground regions or suffer from errors in dealing with the image backgrounds (see Fig 1 (c)).

To overcome the feature synthesizing problems toward real-world detection scenarios, we build a novel zero-shot object detection framework as shown in Fig 2. Specifically, we design two components for learning robust region features. To enable the model to synthesize diverse visual features, we propose an Intra-class Semantic Diverging (IntraSD) component which can diverge the semantic vector of a single class into a set of visual features. To prevent the intra-class diversity of the synthesized features goes too far to mix up the inter-class relationship, we further propose an Inter-class Structure Preserving (InterSP) component that utilizes real visual samples from different object categories to constrain the separability of the synthesized visual features.

It is also worth mentioning that in the design of InterSP, we fully leverage the region features sampled from the real image scenes for detection instead of implementing it on the synthesized visual features. This enables our model to synthesize visual features as separable as in real cases and obtain much better performance when compared to the aforementioned counterpart (see experiments in Section 4.2).

To sum up, this paper mainly has the following three-fold contributions:

- We reveal the key challenges, i.e., the intra-class diversity and inter-class separability, for feature synthesizing in real-world object detection scenarios.
- With the goal to synthesize robust region features for ZSD, we build a novel framework that contains an Intra-class Semantic Diverging component and an Inter-class Structure Preserving component.
- Comprehensive experiments on three datasets, including PASCAL VOC, COCO, and DIOR, demonstrate the effectiveness of the proposed approach. Notably, this is also the first attempt for implementing zero-shot object detection in remote sensing imagery.

2. Related Work

Zero-shot Learning (ZSL). ZSL aims to use seen examples to train the network and reason about unseen classes by leveraging the semantic label embeddings (e.g. word-vector [30] or semantic attributes [2]) as side information [15, 39]. Earlier ZSL research works focus on the embedding function-based methods, which embed the visual features into the semantic descriptor space, or vice versa [1, 5, 12, 21]. As a result, the visual features and the semantic features will lie in a same embedding space and the ZSL classification can be accomplished by searching the nearest semantic descriptor in the embedding space [16]. Embedding function-based methods work well in conventional ZSL scenario [1, 5, 13, 42] but tend to be highly overfitting the seen classes in the more challenging GZSL scenario [7, 16, 37, 43]. To tackle this overfitting problem,

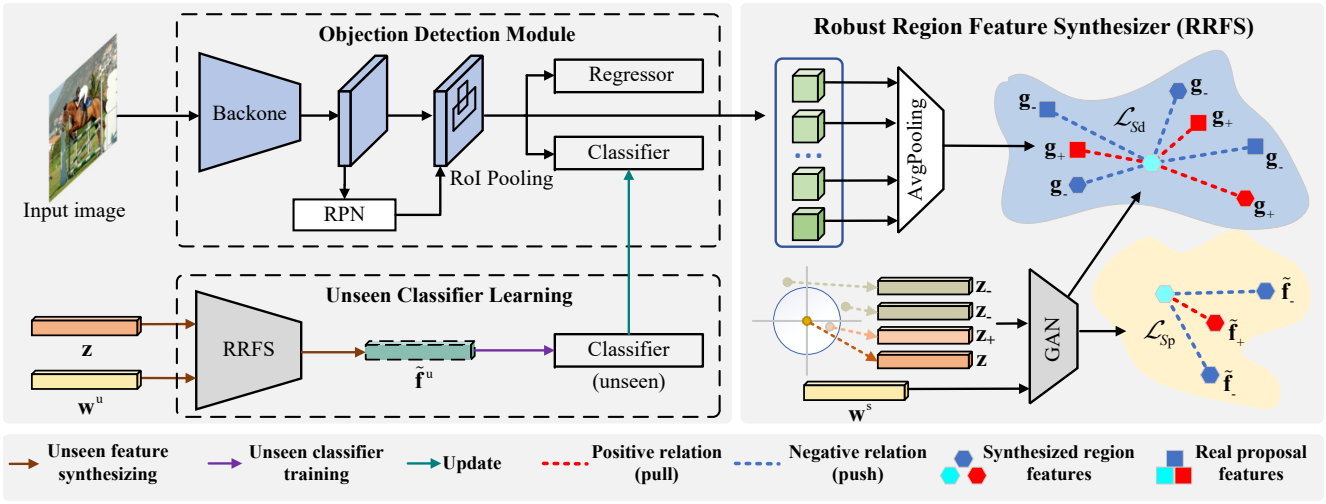


Figure 2. Illustration of the proposed overall framework. Our method contains an object detection module and a unseen classifier learning module. The basic idea is to learn object detector based on the labeled seen category data firstly, and then use the synthesized unseen region features to train unseen classifiers. To keep the framework simple and easy to understand, we do not show the discriminator used in the learning process.

some researchers have introduced generative-based methods [7, 16, 20, 35], which learn to complement the training samples for unseen classes by using a conditional generative model e.g. Variational Autoencoder (VAE) [20] and Generative Adversarial Networks (GAN) [44]. With the synthesized unseen classes examples, they can transform the zero-shot classification problem to a general fully-supervised problem and relieve the overfitting problem [47]. In this paper, we also employ a generative model to synthesize unseen visual features for converting the ZSL into a fully supervised way [38]. However, as our goal is to solve the more challenging ZSD problem, we need to handle heavier intra-class diversity and inter-class separability in model design.

Zero-shot object detection. ZSD receives great research interest in recent years [4, 8, 17, 23, 32, 33, 49–51]. Some researches focus on embedding function-based methods [4,8,23,32,33,50]. Unfortunately, these methods would suffer from the overfitting problem like in ZSL, where the unseen objects are significantly biased towards the seen classes or background [11, 17, 46]. Generative model-based methods [17,49,51] show strong performance in solving the bias problem [33, 40, 45]. Zhu *et al.* [51] synthesized visual features for unseen objects from semantic information and augments existing training algorithms to incorporate unseen object detection. Zhao *et al.* [49] proposed a Generative Transfer Network for zero-shot object detection. Hayat *et al.* [17] proposed a feature synthesis approach for zero-shot object detection and used the mode seeking regularization [28] to enhance the diversity of synthesized features. However, these methods do not have sufficient learning capacity for synthesizing region features that are as intra-class diverse and inter-class separable as the real samples, and this is the core problem studied in this work.

3. Method

3.1. Problem Definition and Framework Overview

In ZSD, we have two disjoint sets of classes: seen classes in $\mathcal{Y}^{\mathrm{s}}$ and unseen classes in $\mathcal{Y}^{\mathrm{u}}$, where $\mathcal{Y}^{\mathrm{s}} \cap \mathcal{Y}^{\mathrm{u}} = \varnothing$. The training set contains seen objects, where each image is provided with the corresponding class labels and bounding box coordinates. In contrast, the test set may contain unseen objects only (i.e., the ZSD setting) or both seen and unseen objects (i.e., the GZSD setting). During learning and testing, the semantic word-vectors $\mathcal{W} = \{\mathcal{W}^{\mathrm{s}}, \mathcal{W}^{\mathrm{u}}\}$ are provided for both seen and unseen classes. The task of ZSD is to learn detectors (parameterized by θ) that can localize and recognize unseen objects corresponding to semantic word-vectors.

Figure 2 shows the proposed overall framework for ZSD. As can be seen, it contains an object detection module and an unseen classifier learning module. The object detection module is a Faster-RCNN model [34] with the ResNet-101 as the backbone [18]. First of all, we train the Faster-RCNN model with seen images and their corresponding ground-truth annotations. Once the model is obtained, we can use it to extract region features using RPN for seen classes. Second, we train the region feature synthesizer to learn the mapping between semantic word-vectors and the visual features. Then, we use the learned feature synthesizer to generate region features for unseen classes. With these synthesized unseen region features and their corresponding class labels, we can train the unseen classifier for unseen classes. Finally, we update the classifier in the Faster-RCNN model to achieve a new detector for the ZSD task. The overall

training procedure is also elaborated in Algorithm 1.

Notice that the core of our method is how to learn a unified generative model to learn the relationship between visual and semantic domains. Specifically, we design a unified region feature synthesizer for feature synthesizing in real-world detection scenarios, which contains an intra-class semantic diverging component and an inter-class structure preserving component.

3.2. The Robust Region Feature Synthesizer

Given the object feature collection $\mathcal{F}^s$, the corresponding label collection $\mathcal{Y}^s$, and the semantic vector $\mathcal{W}^s$ for seen training data $\mathcal{X}^s$, the goal is to learn a conditional generator $G : \mathcal{W} \times \mathcal{Z} \mapsto \mathcal{F}$. That is to say, when we take a class embedding $\mathbf{w} \in \mathcal{W}$ and a random noise vector $\mathbf{z} \sim \mathcal{N}(0,1) \in \mathbb{R}^d$ sampled from a Gaussian distribution as inputs, we can generate the visual feature $\tilde{\mathbf{f}} \in \mathcal{F}$ for the object regions belonging to this class. Then, with the synthesized region features for the unseen classes, we can learn the classifiers for unseen objects. In other words, the generator G learns the mapping between the semantic vectors and the corresponding region features. To learn such a region feature synthesizer, we propose the following learning objective function:

$$\min_G \max_D \mathcal{L}_{\mathrm{WGAN}} + \lambda_1 \mathcal{L}_{\mathrm{C_s}} + \lambda_2 \mathcal{L}_{\mathrm{S_d}} + \lambda_3 \mathcal{L}_{\mathrm{S_p}}, \qquad (1)$$

where $\mathcal{L}_{WGAN}$ is the Wasserstein GAN loss [3] used to enforce the generator to synthesize region features that are aligned well with the distribution of the real region features:

$$\begin{aligned} \mathcal{L}_{\mathrm{WGAN}} = \mathbb{E}[D(\mathbf{f}^{\mathrm{s}}, \mathbf{w}^{\mathrm{s}})] - \mathbb{E}[D(\tilde{\mathbf{f}}^{\mathrm{s}}, \mathbf{w}^{\mathrm{s}})] \\ -\lambda \mathbb{E}[(\| \bigtriangledown_{\hat{\mathbf{f}}^{\mathrm{s}}} D(\hat{\mathbf{f}}^{\mathrm{s}}, \mathbf{w}^{\mathrm{s}})\|_2 - 1)^2], \end{aligned} \qquad (2)$$

where $\mathbf{f}^s$ is the real visual features of the object regions from the seen classes, $\tilde{\mathbf{f}}^{\mathrm{s}} = G(\mathbf{w}^s, \mathbf{z})$ denotes the generated visual features conditioned on class semantic vector $\mathbf{w}^s \in \mathcal{W}^s$, $\hat{\mathbf{f}}^{\mathrm{s}} = \mu \mathbf{f}^{\mathrm{s}} + (1\text{-}\mu)\tilde{\mathbf{f}}^{\mathrm{s}}$ with μ sampled from the uniform distribution $\mu \sim \mathcal{N}(0,1)$, and λ is the penalty coefficient. The discriminator D: $\mathcal{F}^{\mathrm{s}}$ ($\tilde{\mathcal{F}}^{\mathrm{s}}$) $\times \mathcal{W}^{\mathrm{s}} \rightarrow [0,1]$, takes a real region feature $\mathbf{f}^{\mathrm{s}} \in \mathcal{F}^{\mathrm{s}}$ or a synthetic visual feature $\tilde{\mathbf{f}}^{\mathrm{s}}$ with the corresponding class-semantic embedding $\mathbf{w}^{\mathrm{s}}$ as input. It tries to accurately distinguish real visual features from synthetic visual features. In $\mathcal{L}_{\mathrm{WGAN}}$, the first two terms calculate Wasserstein distance, while the third term constrains the gradient of the discriminator G to have unit norm along with the line connecting pairs of real feature $\mathbf{f}^{\mathrm{s}}$ and synthesized feature $\tilde{\mathbf{f}}^{\mathrm{s}}$. $\mathcal{L}_{C_s}$ ensures the generated visual features aligned with the pre-trained classifier on the seen data, which refers to [17].

To improve the robustness of the region feature synthesizer, we explore two new learning terms, including $\mathcal{L}_{S_d}$ and $\mathcal{L}_{S_p}$. Specifically, $\mathcal{L}_{S_d}$ is the proposed intra-class semantic diverging loss, which diverges the semantic word-vector of one object category into a set of region visual features. $\mathcal{L}_{S_p}$ is the proposed Inter-class Structure Preserving loss, whose goal is to constrain the separability of the synthesized visual features. λ_1, λ_2 and λ_3 are the weighting hyper-parameters to balance each component.

Algorithm 1 Training procedure for our framework.

Input: Training image collection with the corresponding class and bounding box annotations $\{\mathcal{X}^{\mathrm{s}}, \mathcal{Y}^{\mathrm{s}}, \mathcal{B}^{\mathrm{s}}\}$, Semantic-word vector collection $\mathcal{W}$;

Output: Object detector parameters $\theta = \{\theta^p, \theta^s, \theta^u\}$ (θ^p indicates the parameter for RPN proposal extraction);

1: $\{\theta^p, \theta^s\} \leftarrow$ Train object detector on $\{\mathcal{X}^{\mathrm{s}}, \mathcal{Y}^{\mathrm{s}}, \mathcal{B}^{\mathrm{s}}\}$;
2: $\mathcal{F}^{\mathrm{s}} \leftarrow$ Extract region features from $\mathcal{X}^{\mathrm{s}}$ using RPN;
3: $G \leftarrow$ Train region feature synthesizer on $\mathcal{F}^{\mathrm{s}}$ and $\mathcal{Y}^{\mathrm{s}}$ by optimizing the loss function in Eq. 1;
4: $\tilde{\mathcal{F}}^{\mathrm{u}} \leftarrow$ Synthesize region features for unseen classes using the trained G and $\mathcal{W}^u$;
5: $\theta^{\mathrm{u}} \leftarrow$ Train unseen object classifier θ^{u} using $\tilde{\mathcal{F}}^{\mathrm{u}}$, $\mathcal{Y}^{\mathrm{u}}$;
6: $\theta \leftarrow$ Update classifier θ of the object detection module with θ^{u};
7: **return** θ ;

3.3. Intra-class Semantic Diverging

To enable the model to synthesize diverse visual features, we consider the IntraSD component to diverge the semantic vector of one semantic word-vector into a set of visual features. Specifically, we conjecture that the above issue could be alleviated by enhancing the influence of noise vectors on the synthetic visual features while preserving the fitness of the synthetic visual features to the class-semantic embeddings. To this end, we propose a novel Intra-class Semantic Diverging loss, where the visual features synthesized from adjacent noise vectors will be pulled closer while those synthesized from distinct noise vectors will be pushed away.

The key design underlying the Intra-class Semantic Diverging loss is how to select the "positive" and "negative" sample pairs. We design positive samples and negative samples by manipulating the input noise vectors [25]. Specifically, given a query noise vector $\mathbf{z} \sim \mathcal{N}(0,1)$, we define a small hyper-sphere with radius r centered at the query noise vector $\mathbf{z}$. We random sample a positive query noise vector $\mathbf{z}_+$ as a vector with the sphere $\mathbf{z}_+ = \mathbf{z} + \boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is a randomly sampled vector from a uniform distribution $\boldsymbol{\rho} \sim \mathcal{U}[-r, r]$. We sample the negative noise vectors as random vectors outside the sphere within a latent space, i.e., $\mathbf{z}_{i-} \sim \{\mathbf{z}_{i-} | \mathbf{z}_{i-} \sim \mathcal{N}(0,1) \cap |\mathbf{z}_{i-} - \mathbf{z}| \succ r\}$ for $i = 1, \ldots, N$, where $\succ$ is the element-wise greater-than operator. Once these noise vectors are determined, we can define the "positive" and "negative" samples. For a query visual feature $\tilde{\mathbf{f}}^{\mathrm{s}} = G(\mathbf{z}, \mathbf{w}^{\mathrm{s}})$ synthesized from the noise vector $\mathbf{z}$, we define

its "positive" sample as $\mathbf{f}^{\mathrm{s}}_{+} = G(\mathbf{z}_{+}, \mathbf{w}^{\mathrm{s}})$ synthesized from the noise vector $\mathbf{z}^{+}$. The N "negative" samples synthesized from the sets of noise vectors $\{\mathbf{z}_{i-}\}$ can be defined as $\mathbf{f}^{\mathrm{s}}_{i-} = G(\mathbf{z}_{i-}, \mathbf{w}^{\mathrm{s}})$. The Intra-class Semantic Diverging loss is given by

$$\mathcal{L}_{\mathrm{S_d}} = \mathbb{E}[-\log \frac{\exp(\tilde{\mathbf{f}}^{\mathrm{s}} \cdot \tilde{\mathbf{f}}^{\mathrm{s}}_{+}/\tau)}{\exp(\tilde{\mathbf{f}}^{\mathrm{s}} \cdot \tilde{\mathbf{f}}^{\mathrm{s}}_{+}/\tau) + \sum_{i=1}^{N} \exp(\tilde{\mathbf{f}}^{\mathrm{s}} \cdot \tilde{\mathbf{f}}^{\mathrm{s}}_{i-}/\tau)}], \tag{3}$$

where "$\cdot$" is the dot product between two visual feature vectors to measure the cosine similarity and τ is a temperature scale factor.

3.4. Inter-class Structure Preserving

In order to make the synthesized visual features approximate distributions of the real data, meanwhile improving the discrimination of the learned visual features, we further introduce the Inter-class Structure Preserving component into the learning framework. In this learning component, we not only consider the synthesized visual features of different categories, but also pay attention to the real region features extracted by the window proposals, which contain both the positive object proposals, i.e., proposals with the same class as the synthesized feature, and many negative and background proposals.

By doing so, the proposed Inter-class Structure Preserving component has the following merits: 1) The proposed method surpasses the reconstruction error-based loss in the conventional WGAN as it can force the synthesized visual features to be close to other different real visual features of the same category in the window proposal pool. By this way, the synthesized visual features can well approximate the distribution of the real data, facilitating robust one-to-many projection from the semantic word vector to the synthesized region features. 2) By pushing away the visual features from different categories (both in real and synthesized feature space), this learning component can effectively enhance the discrimination of the synthesized visual features.

From the above description, we can observe that the proposed method uses both the synthesized region features and real proposal features to implement the learning process, which essentially constructs a hybrid visual feature pool denoted as $\mathbf{g} = \{\tilde{\mathbf{f}}^{\mathrm{s}}, \mathbf{f}^{\mathrm{r}}, \mathbf{f}^{\mathrm{bg}}\}$, where $\mathbf{f}^{\mathrm{r}}$ denotes the real features of the window proposals for different object categories and $\mathbf{f}^{\mathrm{bg}}$ indicates the background visual features extracted from the training images. Then, the learning objective function of the proposed Inter-class Structure Preserving component can be written as:

$$\mathcal{L}_{\mathrm{S_p}} = \mathbb{E}[-\log \frac{\exp(\tilde{\mathbf{f}}^{\mathrm{s}} \cdot \mathbf{g}_{+}/\tau)}{\exp(\tilde{\mathbf{f}}^{\mathrm{s}} \cdot \mathbf{g}_{+}/\tau) + \sum_{j \in \Phi} \exp(\tilde{\mathbf{f}}^{\mathrm{s}} \cdot \mathbf{g}_{\mathrm{j}}/\tau)}], \tag{4}$$

Table 1. Comparison of mAP at IoU=0.5, under ZSD and GZSD settings on PASCAL VOC dataset.

Method	ZSD	GZSD		
		S	U	HM
SAN [33]	59.1	48.0	37.0	41.8
HRE [8]	54.2	**62.4**	25.5	36.2
PL [32]	62.1	-	-	-
BLC [50]	55.2	58.2	22.9	32.9
SU [17]	64.9	-	-	-
Ours	**65.5**	47.1	**49.1**	**48.1**

where $\Phi = \{\mathbf{g}_{\mathrm{j}}\}$ indicates the collection of visual features satisfying $y(\mathbf{g}_j) \neq y(\tilde{\mathbf{f}}^{\mathrm{s}})$ in the hybrid visual feature pool, $y(\cdot)$ is the category indicator function, i.e., $y(\tilde{\mathbf{f}}^{\mathrm{s}})$ is the class label for visual feature $\tilde{\mathbf{f}}^{\mathrm{s}}$. $\mathbf{g}_{+}$ is the positive examples corresponding to the current synthesized visual feature $\tilde{\mathbf{f}}^{\mathrm{s}}$. It can be selected from the synthesized visual features or the object proposals generated by the detector which share the same category label with the current synthesized visual feature $\tilde{\mathbf{f}}^{\mathrm{s}}$. Therefore, this Inter-class Structure Preserving loss enables the synthesized visual feature $\tilde{\mathbf{f}}^{\mathrm{s}}$ to be close to both the synthesized and real object proposals of the same category, while far apart from all other visual features from different class labels in the hybrid visual feature pool.

4. Experiment

Datasets: We evaluate the proposed method on three popular object detection benchmark datasets: PASCAL VOC 2007+2012 [10], MS COCO 2014 [24], and DIOR [22]. The PASCAL VOC 2007 contains 2501 training images, 2510 validation images, and 5011 test images with 20 categories. PASCAL VOC 2012 contains 5717 training images and 5823 validation images also with 20 categories. MS COCO 2014 contains 82783 training images and 40504 validation images with 80 categories. DIOR contains 5862 training images, 5863 validation images, and 11738 test images with 20 categories. For PASCAL VOC and MS COCO, we adopt the FastText method [29] to extract the semantic word-vector following [17]. For DIOR, we adopt the Bert model [9] to generate the semantic word-vector.

Seen/unseen split: We follow the 16/4 seen/unseen split proposed in [8] on the PASCAL VOC dataset. For MS COCO, we adopt the same setting as [4, 32] to divide the dataset with two different splits: (1) 48/17 seen/unseen split (2) 65/15 seen/unseen split. We divide the DIOR dataset with 16/4 seen/unseen split and the details of the split are provided in the supplementary material. For all the above datasets and splits, we remove all the images of unseen categories from the training set to guarantee that unseen objects will not be available during model training.

Evaluation Protocols: We follow the evaluation strat-

Table 2. Class-wise AP and mAP comparison of different methods on unseen classes of PASCAL VOC dataset for ZSD.

Method	car	dog	sofa	train	mAP
SAN [33]	56.2	85.3	**62.6**	26.4	57.6
HRE [8]	55.0	82.0	55.0	26.0	54.5
PL [32]	63.7	87.2	53.2	44.1	62.1
BLC [50]	43.7	86.0	60.8	30.1	55.2
SU [17]	59.6	92.7	62.3	45.2	64.9
Ours	**60.1**	**93.0**	59.7	**49.1**	**65.5**

Table 3. ZSD performance of Recall@100 and mAP with different IoU thresholds on MS COCO dataset.

Method	Split	Recall@100			mAP
		IoU=0.4	IoU=0.5	IoU=0.6	IoU=0.5
SB [4]	48/17	34.5	22.1	11.3	0.3
DSES [4]	48/17	40.2	27.2	13.6	0.5
TD [23]	48/17	45.5	34.3	18.1	-
PL [32]	48/17	-	43.5	-	10.1
BLC [50]	48/17	51.3	48.8	45.0	10.6
Ours	48/17	**58.1**	**53.5**	**47.9**	**13.4**
PL [32]	65/15	-	37.7	-	12.4
BLC [50]	65/15	57.2	54.7	51.2	14.7
SU [17]	65/15	54.4	54.0	47.0	19.0
Ours	65/15	**65.3**	**62.3**	**55.9**	**19.8**

egy proposed in [4, 8]. For PASCAL VOC and DIOR, we utilize mean average precision (mAP) with IoU threshold 0.5 to evaluate the performance. For MS COCO, we utilize mAP with IoU threshold 0.5 and recall@100 with three different IoU thresholds (*i.e.* 0.4, 0.5, and 0.6) as the metric. Furthermore, since the test set consists of seen and unseen images, the performance of GZSD is evaluated by the Harmonic Mean (HM) [15].

Implementation Details: Our object detection module adopts the widely-used Faster-RCNN model [34] with the ResNet-101 as the backbone [18]. The generator G and discriminator D are both two fully-connected layers with LeakyReLU activation [27]. For each unseen class, we synthesize 250/250/500 region features for COCO/DIOR/PASCAL VOC to train the classifiers. The hyperparameter λ_1 in Eq. (1) is set to 0.001/0.001/0.01 for COCO/DIOR/PASCAL VOC. Empirically, in the IntraSD component, the trade-off parameter λ_2 is set to 0.001, the number of negative samples N is set to 10, the temperature coefficient τ is set to 0.1, and the radius r is set to $10^{-4}/10^{-4}/10^{-6}$ for COCO/DIOR/PASCAL VOC. For the IntraSP component, the trade-off parameter λ_3 is set to 0.001 and the temperature coefficient τ is set to 0.1. Code is available at https://github.com/HPL123/RRFS.

Table 4. Comparison of Recall@100 and mAP at IoU=0.5 over two seen/unseen splits, under GZSD setting on MS COCO dataset.

Method	Split	Recall@100			mAP		
		S	U	HM	S	U	HM
PL [32]	48/17	38.2	26.3	31.2	35.9	4.1	7.4
BLC [50]	48/17	57.6	46.4	51.4	42.1	4.5	8.2
Ours	48/17	**59.7**	**58.8**	**59.2**	**42.3**	**13.4**	**20.4**
PL [32]	65/15	36.4	37.2	36.8	34.1	12.4	18.2
BLC [50]	65/15	56.4	51.7	53.9	36.0	13.1	19.2
SU [17]	65/15	57.7	53.9	55.8	36.9	19.0	25.1
Ours	65/15	**58.6**	**61.8**	**60.2**	**37.4**	**19.8**	**26.0**

4.1. Comparison with the State-of-the-art

In Table 1, we compare with state-of-the-art methods on the PASCAL VOC dataset on ZSD and GZSD settings. We can observe that our method outperforms all the comparison methods in terms of the ZSD setting. Compared with the second-best method SU [17], our method improves the mAP of ZSD performance from 64.9 % to 65.5 %. Our method achieves the best performance on unseen classes denoted as "U" among all the comparing methods in terms of GZSD setting. Although our seen performance denoted as "S" is lower, our method achieves the best performance in terms of "HM", which reveals that our method maintains a good balance between seen and unseen classes. This benefits from the robust region feature synthesizer trained with the Intra-class Semantic Diverging component and the Inter-class Structure Preserving component. We also report the class-wise mAP performance in terms of ZSD setting on the PASCAL VOC dataset in Table 2. Our method achieves the best performance on 3 out of 4 classes, which further demonstrates the superiority of our method on ZSD.

In Table 3, we compare our method with the state-of-the-art methods on MS COCO dataset over two splits. For the 47/17 split, our method outperforms all the compared methods by a large margin in terms of both Recall@100 and mAP measurements. Compared with the second-best method BLC [50], our method improves the Recall@100 by 9.6 % and mAP by 26.4 % at IoU=0.5. For the 65/15 split, we can observe that our method also achieves a significant performance gain, which improves the Recall@100 and mAP of method SU [17] from 54.0 % and 19.0 % to 62.3 % and 19.8 % at IoU=0.5.

In Table 4, we compare our method with other methods under the GZSD scenario, which is more realistic and challenging. Our method outperforms all the comparison methods over two splits in terms of all the metrics. Compared with the second-best method BLC [50], our method improves the "HM" performance in Recall@100 and mAP from 51.4 % and 8.2 % to 59.2 % and 20.4 % under the split 48/17. For the 65/15 split, our method improves the "HM"

Table 5. Class-wise AP and mAP comparison of different methods on unseen classes of MS COCO dataset for ZSD.

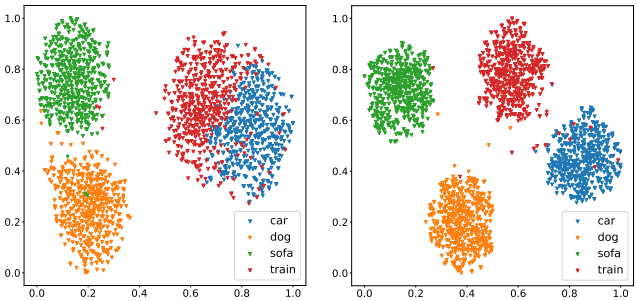
65/15	airp	trai	metr	cat	bear	scse	frbe	snrd	fork	swic	hdog	tlet	mose	tstr	hier	mAP
PL [32]	20	48.2	0.6	28.3	13.8	**12.4**	**21.8**	15.1	8.9	8.5	**0.9**	5.7	0.0	**1.7**	0.0	12.4
SU [17]	10.1	48.7	1.2	64.0	**64.1**	12.2	0.7	28	16.4	19.4	0.1	**18.7**	1.2	0.5	0.2	19.0
Ours	**20.8**	**53.0**	**1.3**	**64.3**	55.5	11.6	0.4	**31.3**	**18.0**	**20.3**	0.1	15.2	**4.2**	0.5	**0.6**	**19.8**

Table 6. Comparison of mAP at IoU=0.5, under ZSD and GZSD settings on DIOR dataset.

Method	ZSD	GZSD		
		S	U	HM
PL [32]	0.4	4.3	0	0
BLC [50]	1.1	6.1	0.4	0.8
SU [17]	10.5	**30.9**	2.9	5.3
Ours	**11.3**	**30.9**	**3.4**	**6.1**

Table 7. Performance of ablation studies under the ZSD and GZSD settings, measured by the mAP on PASCAL VOC dataset.

$\mathcal{L}_{\mathrm{b}}$	$\mathcal{L}_{\mathrm{S_d}}$	$\mathcal{L}_{\mathrm{S_{ps}}}$	$\mathcal{L}_{\mathrm{S_p}}$	ZSD	GZSD		
					S	U	HM
✓				62.1	47.1	45.9	46.5
✓	✓			64.0	47.1	48.3	47.7
✓	✓	✓		64.7	47.1	48.7	47.9
✓	✓		✓	**65.5**	**47.1**	**49.1**	**48.1**



(a) Baseline (b) Ours

Figure 3. The t-SNE [36] visualization of the synthesized region features for the unseen classes on PASCAL VOC dataset.

performance achieved by the second-best method SU [17] from 55.8 % and 25.1 % to 60.2 % and 26.0 %. This "HM" performance gain proves that our region feature synthesizer can synthesize robust features for unseen classes.

We report clas-wise AP of our method in Table 5 for 65/15 split. Our method achieves the best performance on 9 out of 15 classes and comparable performance on other classes. Since other methods did not report their class-wise AP results on the 48/17 split, we show our class-wise AP in the supplementary material, individually.

To further verify the effectiveness of our method, we conduct the experiment on the DIOR dataset, which is the first attempt for implementing zero-shot object detection in remote sensing imagery. We re-implement the state-of-the-art zero-shot object detection methods based on their released codes on the DIOR dataset for comparison in Table 6. Compared with the second-best method SU [17], the "ZSD", "U", and "HM" are improved from 10.5 %, 2.9 %, and 5.3 % to 11.3 %, 3.4 %, and 6.1 %, respectively. We also achieve the same S performance as method SU [17]. Due to space limitation, the concrete class-wise AP scores will be reported in the supplementary material.

4.2. Ablation Study

To provide further insight into our method, we conduct ablation studies on the PASCAL VOC dataset to analyze the contributions of each component in our method. In Table 7, We report the ZSD and GZSD performance in terms of mAP metric at IoU 0.5. $\mathcal{L}_{\mathrm{b}}$ contains the $\mathcal{L}_{\mathrm{WGAN}}$ and $\mathcal{L}_{\mathrm{C_s}}$, which can be regarded as our baseline method. $\mathcal{L}_{\mathrm{S_{ps}}}$ means the hybrid visual feature pool g only contains the synthesized visual features $\tilde{\mathbf{f}}_{\mathrm{s}}$. "✓" denotes the model with the corresponding component.

IntraSD Component Analysis. We first analyze the effectiveness of the proposed IntraSD component. To verify its contributions, we compare the performances of our baseline model and the variant by adding the IntraSD during training. We can observe that the "ZSD" performance and the "U" performance of GZSD all have been improved significantly from 62.1 % and 45.9 % to 64.0 % and 48.3 %, respectively. The HM performance is improved from 46.5 % to 47.7 %. These large performance gains demonstrate the effectiveness of the proposed IntraSD component in the ZSD model, which can encourage our generator to synthesize more diverse visual features for unseen classes. The seen performances of GZSD do not obtain performance gains since the parameters of the classifier for seen class is fixed.

InterSP Component Analysis. To verify the effectiveness of the InterSP component, we first add the $\mathcal{L}_{\mathrm{S_{ps}}}$ component on top of the $\mathcal{L}_{\mathrm{b}}$ and $\mathcal{L}_{\mathrm{S_d}}$. As a result, the mAP measurements of "ZSD", "U", and "HM" are improved from 64.0 %, 48.3 %, and 47.7 % to 64.7 %, 48.7 %, and 47.9 %. Secondly, we further add the $\mathcal{L}_{\mathrm{S_p}}$ component on top of the $\mathcal{L}_{\mathrm{b}}$ and $\mathcal{L}_{\mathrm{S_d}}$. The mAP measurement of "ZSD", "U", and "HM" are improved from 64.0 %, 48.3 %, and 47.7 % to 65.5 %, 49.1 %, and 48.1 %. These two comparisons demonstrate that our InterSP component can improve

Figure 4. Qualitative results on PASCAL VOC, MS COCO (48/17 and 65/15) and DIOR datasets. For each dataset, the first column and second column are the results of ZSD and GZSD, respectively. Seen classes are shown with green and unseen with red.

the discrimination of the learned visual features. Compared with the method variant $\mathcal{L}_{\mathrm{S_{ps}}}$, the variant with $\mathcal{L}_{\mathrm{S_p}}$ gains an absolute improvement of 1.2 %, 0.8 %, and 0.4 %. This phenomenon indicates that the real features of the window proposals play an important role in our InterSP module, since it contains the ground-truth positive object proposals and many background negative proposals.

4.3. Qualitative Results

Feature visualization In Fig 3, we conduct the t-SNE [36] visualization for the synthesized region features of the unseen classes on the PASCAL VOC dataset. The visual feature distributions corresponding to our baseline model and the proposed model have been illustrated in Fig 3(a) and Fig 3(b). Features from similar classes (car and train in Fig 3(a)) are confused with each other due to high similarity in their semantic space, which may lead to miss-classify these similar classes. The synthesized features in Fig 3(b) have obvious segregated clusters. This verifies that our synthesizer is robust in synthesizing intra-class diverse and inter-class discriminative region features, which benefits learning a more discriminative classifier to improve the detection performance for ZSD.

Detection Results. To further show the effectiveness of our method, we show the detection results of our method on PASCAL VOC, MS COCO, and DIOR datasets in Fig 4. For the ZSD setting, the images only contain unseen objects. For the GZSD setting, the images may contain seen and unseen objects together. The qualitative results prove the effectiveness of our method in detecting seen and unseen objects simultaneously in the challenging scenarios.

5. Conclusion and Limitation

In this work, we focus on ZSD task by addressing the challenge of synthesizing robust region features for unseen objects. Specifically, we propose a novel ZSD framework by constructing a robust region feature synthesizer, which includes the IntraSD and InterSP components. The IntraSD realizes the one-to-more mapping to obtain diverse visual features from each class semantic vector, preventing miss-classifying the real unseen objects as image backgrounds. The InterSP component improves the discrimination of the synthesized visual features by make full use of both synthesized and real region features from different object categories. Extensive experimental results demonstrate that our method is superior to state-of-the-art approaches for ZSD.

Limitation. One major limitation in this study is that the proposed method is based on the two-stage object detector, e.g., Faster-RCNN [34], whose detection speed is relatively slow. We hope to integrate our method into some one-stage object detector, e.g., YOLOv5 [19], to further improve the detection speed in the future.

Acknowledgments: This work was supported in part by Key-Area Research and Development Program of Guangdong Province (No.2021B0101200001), and the National Natural Science Foundation of China under Grant 61876140, 62176198, U20B2065, U21B2048, and the Open Research Projects of Zhejiang Lab (No.2019KD0AD01/010).

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826, 2013. 2

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 1, 2

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017. 4

[4] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, pages 384–400, 2018. 1, 3, 5, 6

[5] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, pages 730–746. Springer, 2016. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[7] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *ICCV*, pages 122–131, 2021. 2, 3

[8] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *BMVC*, 2018. 3, 5, 6

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 5

[11] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *CVPR*, pages 10699–10709, 2020. 3

[12] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, pages 584–599. Springer, 2014. 2

[13] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, pages 2635–2644, 2015. 2

[14] Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100, 2018. 1

[15] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, pages 2371–2381, 2021. 2, 6

[16] Zongyan Han, Zhenyong Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. In *CVPR*, pages 12865–12874, 2020. 2, 3

[17] Nasir Hayat, Munawar Hayat, Shafin Rahman, Salman Khan, Syed Waqas Zamir, and Fahad Shahbaz Khan. Synthesizing the unseen for zero-shot object detection. In *ACCV*, 2020. 1, 2, 3, 4, 5, 6, 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[19] Glenn Jocher, Alex Stoken, Jirka Borovec, Ayush Chaurasia, L Changyu, VA Laughing, A Hogan, J Hajek, L Diaconu, Y Kwon, et al. ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations. *Version v5. 0. Apr*, 2021. 8

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[21] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017. 2

[22] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J PHOTOGRAMM*, 159:296–307, 2020. 5

[23] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *AAAI*, volume 33, pages 8690–8697, 2019. 3, 6

[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 *ECCV*, pages 740–755, 2014. 5

[25] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *CVPR*, pages 16377–16386, 2021. 4

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 1

[27] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013. 6

[28] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, pages 1429–1437, 2019. 3

[29] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *LREC*, 2018. 5

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013. 1, 2

[31] Ayyappa Pambala, Titir Dutta, and Soma Biswas. Generative model with semantic embedding and integrated classifier for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1237–1246, 2020. 1

[32] Shafin Rahman, Salman Khan, and Nick Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018. 3, 5, 6, 7

[33] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, pages 547–563. Springer, 2018. 1, 3, 5, 6

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2016. 1, 3, 6, 8

[35] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019. 3

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7, 8

[37] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, pages 4281–4289, 2018. 2

[38] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhijie Zhang. Gatector: A unified framework for gaze object prediction. In *CVPR*, 2022. 3

[39] Binglu Wang, Xun Zhang, and Yongqiang Zhao. Exploring sub-action granularity for weakly supervised temporal action localization. *CSVT*, 2021. 2

[40] Wenguan Wang, Tianfei Zhou, Siyuan Qi, Jianbing Shen, and Song-Chun Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. *TIP*, 2021. 3

[41] Jiamin Wu, Tianzhu Zhang, Zheng-Jun Zha, Jiebo Luo, Yongdong Zhang, and Feng Wu. Self-supervised domain-aware generative network for generalized zero-shot learning. In *CVPR*, pages 12767–12776, 2020. 1

[42] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016. 2

[43] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 2

[44] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pages 10275–10284, 2019. 3

[45] Le Yang, Junwei Han, and Dingwen Zhang. Colar: Effective and efficient online action detection by consulting exemplars. In *CVPR*, 2022. 3

[46] Le Yang, Junwei Han, Tao Zhao, Tianwei Lin, Dingwen Zhang, and Jianxin Chen. Background-click supervision for temporal action localization. *PAMI*, 2021. 3

[47] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *TIP*, 29:8535–8548, 2020. 3

[48] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. In *ICCV*, pages 1801–1810, 2017. 1

[49] 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 *AAAI*, volume 34, pages 12967–12974, 2020. 2, 3

[50] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In *ACCV*, 2020. 3, 5, 6, 7

[51] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Don't even look once: Synthesizing features for zero-shot detection. In *CVPR*, pages 11693–11702, 2020. 1, 3