# Zero-Shot Aerial Object Detection with Visual Description Regularization

**Zhengqing Zang**[1,2*], **Chenyu Lin**[1,2*], **Chenwei Tang**[1,2], **Tao Wang**[1,2†], **Jiancheng Lv**[1,2]

[1]College of Computer Science, Sichuan University, Chengdu, 610065, P. R. China
[2]Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, 610065, P. R. China
{2022223045158, 2022223040017}@stu.scu.edu.cn, tangchenwei@scu.edu.cn, twangnh@gmail.com
lvjiancheng@scu.edu.cn

## Abstract

Existing object detection models are mainly trained on large-scale labeled datasets. However, annotating data for novel aerial object classes is expensive since it is time-consuming and may require expert knowledge. Thus, it is desirable to study label-efficient object detection methods on aerial images. In this work, we propose a zero-shot method for aerial object detection named visual Description Regularization, or *DescReg*. Concretely, we identify the weak semantic-visual correlation of the aerial objects and aim to address the challenge with prior descriptions of their visual appearance. Instead of directly encoding the descriptions into class embedding space which suffers from the representation gap problem, we propose to infuse the prior inter-class visual similarity conveyed in the descriptions into the embedding learning. The infusion process is accomplished with a newly designed similarity-aware triplet loss which incorporates structured regularization on the representation space. We conduct extensive experiments with three challenging aerial object detection datasets, including DIOR, xView, and DOTA. The results demonstrate that DescReg significantly outperforms the state-of-the-art ZSD methods with complex projection designs and generative frameworks, e.g., DescReg outperforms best reported ZSD method on DIOR by 4.5 mAP on unseen classes and 8.1 in HM. We further show the generalizability of DescReg by integrating it into generative ZSD methods as well as varying the detection architecture. Codes will be released at https://github.com/zq-zang/DescReg.

## Introduction

Aerial object detection aims to detect objects from aerial images (Xia et al. 2018; Yang et al. 2019; Ding et al. 2021), e.g., images captured from an unmanned aerial vehicle (UAV). It plays an important role in many remote sensing applications, such as UAV-aided environmental monitor and disaster response systems. Benefiting from the development of deep convolution neural networks (CNNs), aerial object detection has been extensively studied and advanced (Yang et al. 2019; Zhu et al. 2021; Li et al. 2022, 2020; Deng et al. 2020; Han et al. 2021) in recent years. Prior research mainly focuses on improving the accuracy or efficiency based on a fully supervised paradigm. However, labeling objects for large-scale aerial images is extremely costly due to the small object size and irregular viewing angle. Hence expanding the vocabulary becomes a challenge for fully supervised aerial object detection methods (Lam et al. 2018).

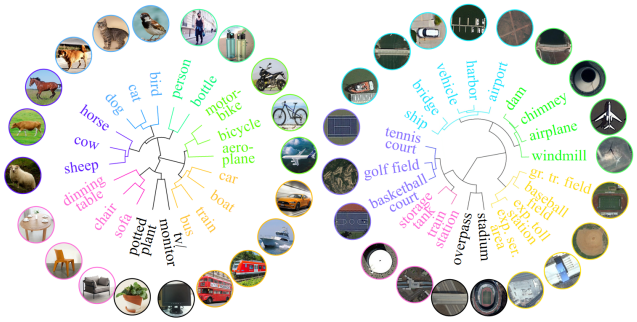


Figure 1: Illustration of weak semantic-visual correlation problem. We perform hierarchical clustering with semantic embeddings and show the radial dendrogram for the 20 common object classes from the Pascal VOC dataset (left) and the 20 aerial object classes from the DIOR dataset(right). The common object classes show clear clustering result which corresponds well to visual appearance (e.g., horse, cow, and sheep), while the semantic clustering of aerial object classes are inevident and shows much less correlation with visual appearance. Best viewed with zoom-in.

*Zero-shot object detection* (ZSD), which aims to detect unseen object classes without bounding box annotations (Bansal et al. 2018; Demirel, Cinbis, and Ikizler-Cinbis 2018; Rahman, Khan, and Porikli 2018), appears as a promising approach for reducing the copious label demand in aerial object detection. ZSD methods mainly leverage the semantic relation between object classes to detect unseen classes, e.g., Cat and Dog are semantically similar, and thus the knowledge learned on Cat could be transferred to recognize Dog. Methodologically, this knowledge transfer process is typically realized through learning a command embedding function to align visual and semantic features (Rahman, Khan, and Porikli 2018; Bansal et al. 2018; Zheng et al. 2020; Rahman, Khan, and Barnes 2020), or learning a universal synthesizer function to generate training sam-

*These authors contributed equally.
†Corresponding author

ples (Hayat et al. 2020; Zhao et al. 2020; Zhu, Wang, and Saligrama 2019; Huang et al. 2022; Sarma, Kumar, and Sur 2022). However, we find that existing ZSD methods perform poorly on aerial images due to weak semantic-visual correlation. Concretely, as shown in Fig. 1, our core observation is that objects in natural images tend to be visually distinct and align well with semantic clustering, yet objects from aerial images often appear vague and lack semantic correlation. Such an issue hinders effective recognition of unseen classes.

Based on the analysis, we aim to incorporate textual descriptions to enhance the semantic understanding of aerial object classes. These descriptions, which detail visual characteristics, act as prior knowledge. We initially encoded these descriptions using semantic embeddings from a pretrained language model, noting a performance improvement, though limited. This limitation is likely due to the visual-semantic representation gap (Wang and Chen 2017), more pronounced in aerial images. Consequently, we shift our approach, using textual descriptions for structural regularization. Our proposed method, Description Regularization (*DescReg*), aims to maintain the visual similarity structure in the classification space, enhancing knowledge transfer from seen to unseen classes. For this, we designed an adaptive triplet loss, treating each projected class embedding as separate samples. This involves sampling positive pairs from similar classes and negative pairs from dissimilar ones, using their difference as the margin. This similarity-aware triplet loss effectively preserves inter-class similarity relations in the embedding space during optimization.

To validate the above method, we establish two challenging zero-shot aerial object detection setups with DOTA and xView datasets. Together with the existing aerial ZSD setup on the DIOR dataset (Huang et al. 2022), we conduct extensive experiments on the two-stage Faster R-CNN detector and further show generalization the multi-stage Cascaded R-CNN detector (Cai and Vasconcelos 2018) and the popular one-stage YOLOv8 detector (Redmon and Farhadi 2017; Jocher, Chaurasia, and Qiu 2023). DescReg effectively improves the detection accuracy of raw baseline method on both seen and unseen classes. Remarkbaly, DescReg with simple one-layer projection outperforms the SOTA generative ZSD methods (Huang et al. 2022) by 4.5 in unseen mAP and 8.1 in HM, with the same detection architecture. We further incorporate our method into the generative ZSD method by regularizing the visual feature synthesizing process and observe significant improvement, which demonstrates the strong generalizability of our DescReg as a structural similarity regularization method.

In summary, Our contributions are four-fold:

- Our study is the first comprehensive analysis in zero-shot aerial object detection, combining thorough investigation with specialized method development.
- Addressing the weak semantic-visual link in aerial imagery, we use prior visual text descriptions as a solution.
- We introduce a novel triplet loss that accounts for inter-class similarity, embedding structural regularization through textual descriptions.
- Utilizing the DOTA and xView datasets, we establish two new challenging ZSD setups and conduct extensive experiments with various detection architectures to assess our method.

# Related Work

## Zero-shot Object Detection

Driven by zero-shot learning (ZSL) research (Mishra et al. 2018; Tang et al. 2019; Jasani and Mazagonwalla 2019; Demirel, Cinbis, and Ikizler-Cinbis 2019; Tang et al. 2020, 2021), which transfers knowledge from seen to unseen classes, the challenging task of zero-shot detection (ZSD) has gained attention since its introduction in 2018 (Bansal et al. 2018). ZSD not only categorizes but also localizes unseen objects. Similar to ZSL, ZSD strategies are either embedding-based or generative-based. Embedding-based methods learn a visual→semantic projection for aligning two spaces (Demirel, Cinbis, and Ikizler-Cinbis 2018; Li et al. 2019b), including refining background vectors for better differentiation from unseen classes (Zheng et al. 2020). Alternatively, generative methods (Zhu, Wang, and Saligrama 2019; Zhao et al. 2020) use GANs to create visual samples of unseen classes, enabling classifier and regressor training. These approaches focus on maintaining inter-class structure and increasing intra-class diversity, with novel components like robust feature synthesizers (Huang et al. 2022) and loss functions (Sarma, Kumar, and Sur 2022) for visual-semantic alignment. However, the critical role of category embedding representation in ZSD's effectiveness, which is our study's focus, remains underexplored.

## Aerial Object Detection

Aerial images, taken by sensors on satellites, aircraft, or drones, are essential for gathering Earth's surface data from afar. While object detection in natural images has advanced significantly, it remains a challenge in aerial imagery. Previous research has mainly addressed aerial-specific issues like small target sizes (Recasens et al. 2018; Yang et al. 2019; Meethal, Granger, and Pedersoli 2023; Li et al. 2020; Yang, Huang, and Wang 2022; Koyun et al. 2022) and object rotation (Cheng, Zhou, and Han 2016; Zhang et al. 2020b; Cheng et al. 2019). However, there's limited focus on label-efficient detection methods in this field. Some have explored few-shot learning (Wolf et al. 2021; Lu et al. 2023) for aerial detection, but these still need target labels. Our work investigates using training data from known classes for direct application to unknown classes, i.e., zero-shot detection.

# Proposed Method

## Overview

Given the bounding box annotation on a set of seen object categories $\mathbb{F} = \{C_1, C_2, ...C_N\}$, zero-shot object detection (ZSD) aims at training on the seen data and generalizing to a set of target unseen object categories $\mathbb{F}^* = \{C_1, C_2, ...C_M\}$. In the following paragraphs, we first present our main detection architecture and then introduce our approach with in-depth analyses under the context of ZSD.

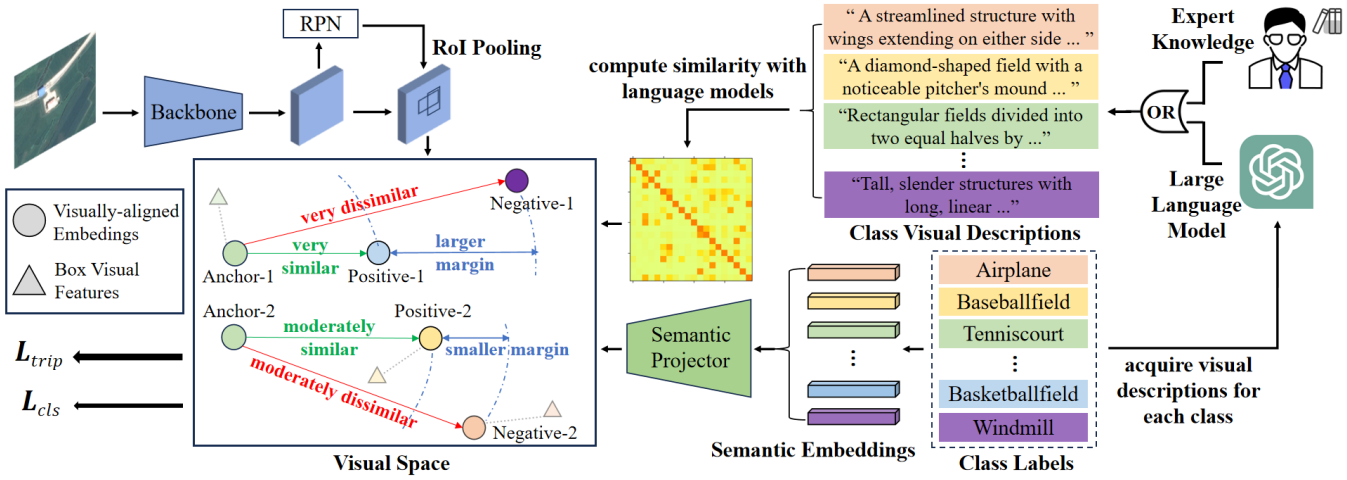


Figure 2: The overall framework of the proposed method.

**Object Detection Architecture** The classical two-stage Faster R-CNN (Ren et al. 2015) detection model consists of a visual feature extraction backbone $\mathcal{T}$, a region proposal network $\mathcal{R}$, a shared multi-layer feature transformation network $\mathcal{F}$. a region classifier $\mathcal{C}$, and a box regressor $\mathcal{B}$. Given input image $\mathbf{I}$, the model first extracts image feature $\mathbf{F}$: $\mathbf{F} = \mathcal{T}(\mathbf{I})$. Then object candidate proposals are predicted by the region proposal network: $\{\mathbf{p}_i\} = \mathcal{R}(\mathbf{F})$. With the proposals, feature pooling is conducted on the image feature map $\mathbf{F}$ to obtain the proposal region feature $\{\mathbf{v}_i\}$. The feature is then further refined by the shared network: $\mathbf{v}_i = \mathcal{F}(\mathbf{f}_i)$. Finally, object classification scores and refined bounding boxes are predicted by the classifier and the regressor: $\mathbf{s}_i = \mathcal{C}(\mathbf{v}_i)$, $\mathbf{b}_i = \mathcal{B}(p_i, \mathbf{v}_i)$.

The detection model can be trained on the seen class data with proposal loss, classification loss, and regression loss: $\mathcal{L} = \mathcal{L}_{prop} + \mathcal{L}_{cls} + \mathcal{L}_{reg}$. The trained region proposal network is class-agnostic and thus may generalize directly to predict the unseen classes. The box regressor is also not sensitive to classes and thus can be applied directly to unseen classes, i.e., by using the class-agnostic version or using prediction from seen classes (Huang et al. 2022). The major challenge here is to generalize the classification to unseen classes, as the region classifier is only trained on the seen class data and cannot predict the unseen classes.

**Detecting the Unseen with Semantic Bridging** While unseen class data is not available, the semantic relation can be efficiently represented with semantic word embeddings. These embeddings can be obtained from pre-trained word embedding models such as Word2vec (Mikolov et al. 2013) and large language models such as BERT (Devlin et al. 2018): $\mathbf{c}_j = \mathcal{W}(C_j)$, where $\mathbf{c}_j$ is the vectorized representation and $\mathcal{W}$ is the embedding model. With these embeddings and trained detection models on the seen class data, existing zero-shot object detection methods mainly focus on bridging the gap between seen and unseen classes. These methods can be classified into *embedding-alignment* and *generative* methods. The embedding-alignment methods (Khandelwal et al. 2023; Zhang et al. 2020a; Yan et al. 2022) aim to bridge the gap between visual and semantic space by learning representation alignment. For example, learning an alignment function $\phi$ to align semantic embeddings to visual features (Zhang et al. 2020a):

$$\mathbf{w}_j = \phi(\mathbf{c}_j) \tag{1}$$

where $w_j$ is the visually-aligned class representation. With $w_j$, the visual features $v_i$ can be classified based on similarity metrics such as cosine similarity, and the seen classification supervision is employed to learn the alignment function. The learned alignment function is expected to generalize to unseen classes by utilizing unseen class embeddings, and thus the detection model can detect unseen objects.

**The Semantic-Visual Correlation Challenge** Although such embedding-alignment methods are shown to be effective on natural image datasets such as Pascal (Everingham et al. 2010) and COCO (Lin et al. 2014). They suffer from poor semantic-visual correlation on aerial images: *the semantic embedding $\mathbf{c}_j$ has poor correlation with visual features $\mathbf{v}_i$, which leads to severe difficulty for the learned embedding function $\phi$ to generalize on the unseen classes.* Note this observation also applies to the generative methods which will be discussed in the next section.

Based on this observation, we aim to improve the semantic-visual correlation by augmenting the semantic embeddings with extra visual cues. The visual cues are instantiated as simple textual descriptions. This is motivated by the prior works in zero-shot learning that employs textual descriptions to augment the recognition of unseen classes (Elhoseiny, Saleh, and Elgammal 2013; Paz-Argaman et al. 2020). The descriptions can be obtained from *experts* or simply through large-scale pre-trained *large language models* (LLM), e.g., GPT (OpenAI 2023).

## DescReg Formulation

The textual descriptions are free-form and efficient to acquire, they can help provide valuable information such as

shape, color, and context for the aerial objects. However, we find simply encoding them into the semantic representation offers limited gain. This is likely due to the following issues: **1)** the representations of semantic feature space and visual feature space have distinct distributions, which causes ineffective transformation of the descriptions into visual feature space. **2)** The image feature representations of aerial objects are less discriminative due to their smaller size than that of common objects, thus it is more difficult to classify them against the false classes and backgrounds.

To address the above issues, we leverage the inter-class visual similarity information as a structural regularization to learn more discriminative alignment functions. Specifically, given the visual descriptions for all seen and unseen classes: $\{T_j\}$, pre-trained language models such as BERT (Devlin et al. 2018) are used to encode them into vectorized representation:

$$\mathbf{t}_j = \mathcal{W}(T_j) \tag{2}$$

where $\mathcal{W}$ is the employed language model and $\mathbf{t}_j$ is the obtained representation. Then we compute the pair-wise cosine similarity and obtain the similarity matrix $\mathbf{S}$:

$$\mathbf{S}(j,k) = \frac{\mathbf{t}_j \mathbf{t}_k}{||\mathbf{t}_j||_2 ||\mathbf{t}_k||_2} \tag{3}$$

To encourage more discriminative inter-class similarity and keep the similarity score within the value range $(0, 1]$, we introduce a self-excluding Softmax:

$$\hat{\mathbf{S}}(j,k) = \begin{cases} \frac{e^{\mathbf{S}(j,k)/\tau}}{\sum_{k' \neq j} e^{\mathbf{S}(j,k')/\tau}} & \text{if } k \neq j \\ \mathbf{S}(j,k) & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{\mathbf{S}}$ is the normalized similarity matrix, of which all the diagonal elements are 1, corresponding to self-similarity and the other elements are in the value range $(0, 1)$, corresponding to inter-class similarity.

The visual characteristics of object classes are now encoded structurally as this similarity matrix. We then integrate it into the embedding alignment learning process. Motivated by the triplet loss (Frome et al. 2013; Akata et al. 2015), we treat the visually-aligned semantic representations $\mathbf{w}_j$ as independent feature samples and perform positive-negative sampling based on the similarity, then triplet loss is imposed on the samples:

$$\mathcal{L}^j_{trip} = max\{0, d(\mathbf{w}_j, \mathbf{w}_{h(j)}) - d(\mathbf{w}_j, \mathbf{w}_{l(j)}) + \Delta\} \tag{5}$$

where $d(\cdot, \cdot)$ is the Euclidean distance. $h(j)$ denotes sampling a similar class for class $j$ and $l(j)$ means sampling a less similar class, or dissimilar class. The sampling is conducted based on similarity scores $\hat{\mathbf{S}}(j, :)$. $\Delta$ is the margin. However, such a direct adoption of triplet loss does not consider the similarity level between classes, e.g., a bridge may be very similar to a dam, but less similar to an overpass, while a vehicle may look a bit dissimilar to a boat but is very distinct to a baseball-field. We thus propose to employ the similarity gap as the margin for the triplet regularization:

$$\Delta_j = \hat{\mathbf{S}}(j, h(j)) - \hat{\mathbf{S}}(j, l(j)) \tag{6}$$

such a second-order metric helps encode the discrepancy in similarity level into the margin regularization. It facilitates the structural learning of the alignment function and thus the knowledge learned from seen classes can better transfer to the unseen classes. The improved similarity-aware triplet loss is thus:

$$\mathcal{L}^j_{trip} = max\{0, d(\mathbf{w}_j, \mathbf{w}_{h(j)}) - d(\mathbf{w}_j, \mathbf{w}_{l(j)}) + \Delta_j\} \tag{7}$$

Unlike prior works that apply contrastive objectives (Huang et al. 2022; Yan et al. 2022), the proposed margin-adaptive triplet loss is less greedy and allows strong flexibility in representation space. The loss is summed over all the seen and unseen classes to compute the full regularization objective:

$$\mathcal{L}_{trip} = \sum_j \mathcal{L}^j_{trip} \tag{8}$$

During the learning of the alignment function, the classification objective (e.g., cross-entropy) is usually computed on the seen classes. So the complete objective with DescReg is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{trip} \tag{9}$$

Fig. 2 shows the overall framework integrated with Faster R-CNN.

**Generalization to Generative Methods** Unlike the above embedding-alignment methods, the generative methods (Hayat et al. 2020; Zhu, Wang, and Saligrama 2019; Huang et al. 2022; Rahman, Khan, and Barnes 2020) aim to learn universal visual feature synthesizers. The method can be simplified as generating visual feature samples based on semantic embeddings:

$$\hat{\mathbf{v}}_j = \varphi(\mathbf{c}_j, \mathbf{z}) \tag{10}$$

where $\hat{\mathbf{v}}_j$ is the synthesized visual feature and $\mathbf{z}$ is random noise to encourage the feature diversity. The synthesized features can be employed to train the classifier for both seen and unseen classes. Similar to the above-mentioned semantic-visual correlation challenge, here the synthesizer also faces generalization issues on the unseen classes. The proposed similarity-aware triplet loss can then easily added to the training process of generative networks:

$$\mathcal{L}^j_{trip} = max\{0, d(\hat{\mathbf{v}}_j, \hat{\mathbf{v}}_{h(j)}) - d(\hat{\mathbf{v}}_j, \hat{\mathbf{v}}_{l(j)} + \Delta_j\} \tag{11}$$

# Experiments

We study four questions in experiments. 1) How does DescReg improve the performance of zero-shot aerial object detection? is it efficient? 2) Is DescReg sensitive to visual descriptions and embedding generation methods? 3) How does each component take effect? 4) Can DescReg generalize to the generative ZSD methods and be applied on different object detection meta-architectures?

## Datasets and Experiment Setup

We evaluate the proposed method on three challenging remote sensing image object detection datasets: DIOR (Li et al. 2019a), xView (Lam et al. 2018), and DOTA (Xia

| Method | ZSD | | | | GZSD | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Recall@100 | | | mAP | Recall@100 | | | mAP | | |
| | IoU=0.4 | IoU=0.5 | IoU=0.6 | | S | U | HM | S | U | HM |
| BLC (Zheng et al. 2020) | - | - | - | - | - | - | - | 6.1 | 0.4 | 0.8 |
| SU (Hayat et al. 2020) | - | - | - | 10.5 | - | - | - | 30.9 | 2.9 | 5.3 |
| RRFS (Huang et al. 2022) | - | - | - | 11.3 | - | - | - | 30.9 | 3.4 | 6.1 |
| V2S† (Khandelwal et al. 2023) | 14.1 | 11.9 | 10.1 | 4.1 | 78.2 | 15.8 | 26.3 | 57.0 | 1.4 | 2.7 |
| RRFS† (Huang et al. 2022) | 22.1 | 19.8 | 18.1 | 9.7 | 60.0 | 19.9 | 29.9 | 41.9 | 2.8 | 5.2 |
| ContrastZSD† (Yan et al. 2022) | 24.9 | 22.3 | 20.1 | 8.7 | 69.2 | 25.9 | 37.7 | 51.4 | 3.9 | 7.2 |
| DescReg (ours) | **37.9** | **34.6** | **31.5** | **15.2** | **82.0** | **34.3** | **48.4** | **68.7** | **7.9** | **14.2** |

Table 1: Comparison with state-of-the-art methods under ZSD and GZSD settings on DIOR dataset. † denotes our implementation results. "S" and "U" denote seen classes and unseen classes, respectively.

et al. 2017). For DIOR, we follow the setting in prior work (Huang et al. 2022). For xView and DOTA, we conduct semantic clustering and sample classes within clusters to ensure unseen class diversity and semantic relatness(Rahman, Khan, and Porikli 2018; Huang et al. 2022). The resulting xView contains 48 seen classes and 12 unseen classes, and the resulting DOTA contains 11 seen classes and 4 unseen classes. We also perform cropping on the xView and DOTA images to simplify the data. Due to space limits, please refer to our supplementary file for more details. Throughout the experiments, unless otherwise stated, we adopt the Faster R-CNN model as the base detection model and IOU=0.5 for the evaluation.

## Implementation Details

Following prior works (Yan et al. 2022; Huang et al. 2022; Yan et al. 2022), we adopt Faster R-CNN with ResNet-101 (He et al. 2016) as the base detection architecture and conduct two-stage training. In the first stage, the model is first trained on the seen class data as conventional detection training, In the second stage, the model is frozen and the semantic-visual projection is fine-tuned with the proposed DescReg. In addition to Faster R-CNN, we also validate our method on the newly released one-stage YOLOv8 model (Jocher, Chaurasia, and Qiu 2023) and the cascaded detection model (Cai and Vasconcelos 2018). Due to space limit, please refer to supplementary for more details on implementation.

## Main Results

**Comparison with State-of-the-arts on DIOR** In Tab. 1, we compare the results with state-of-the-art methods on the DIOR dataset. The proposed method outperforms all compared methods in both ZSD and GZSD settings. *Under the ZSD setting*, our method achieves more than 11.0% absolute gain for recalls of different IOU thresholds, and nearly 4.0% mAP increase compared to the best-reported method, demonstrating its much stronger ability to detect unseen categories compared to All other ZSD methods. *Under the GZSD setting*, the proposed method achieves the best mAP performance on seen classes, surpassing the best-compared method by 11.7% in mAP, this result shows that our zero-shot learning method achieves the least interference on the seen class recognition. Furthermore, our method achieves 7.9% unseen mAP and 14.2% HM, which also significantly outperforms the prior methods. Similar observations hold on the recall metrics.

**Experiments on xView and DOTA** In addition to DIOR, we further conduct zero-shot detection experiments on the challenging xView and DOTA datasets. We compare to RRFS (Huang et al. 2022) and ContrastZSD (Yan et al. 2022) as representatives of SOTA generative methods and embedding-alignment methods. The result is shown in Tab. 2. On both datasets, our method shows higher performances compared to the baselines. Specifically, Under both ZSD and GZSD settings of xView, the proposed method achieves nearly two-fold improvement in unseen mAP compared to the best-performing ContrastZSD method (4.1% to 8.3%, 2.9% to 5.8%). The corresponding gains on DOTA are about 50% relatively. We also observe that with xView, ContrastZSD achieves similar or higher recalls on the GZSD setting compared to our method, but the unseen mAP is lower, which indicates its unseen images may be less discriminative against the background, and thus predicts more false positives.

**Class-wise Resutls** We also report the class-wise mAP performance in terms of ZSD and GZSD for all three aerial object detection datasets. The results are shown in Tab. 3. We note some unseen classes are very challenging and show near 0% AP on the test set (e.g. 0.1% and 0.4% for helicopters on DOTA, under ZSD and GZSD settings respectively). This phenomenon is also observed in prior ZSD works (Yan et al. 2022; Huang et al. 2022; Yan et al. 2022), it is mainly caused by the weak discriminability of unseen class representations and remains a good topic for future ZSD research. Notably, benefiting from the introduced cross-class representation regularization, our method achieves relatively good performances on many unseen classes(e.g. 20.0% GZSD AP and 45.7% ZSD AP for groundtrackfield class on DIOR).

## Generalizability

We further validate whether DescReg generalizes to the generative ZSD method and other detection architectures.

**DescReg with Generative ZSD Methods** Generative methods aim at synthesizing samples for unseen classes, the

| Method | xView | | | | | | | | | | DOTA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ZSD | | | | GZSD | | | | | | ZSD | | | | GZSD | | | | | |
| | RE@100 | | | mAP | RE@100 | | | mAP | | | RE@100 | | | mAP | RE@100 | | | mAP | | |
| | 0.4 | 0.5 | 0.6 | | S | U | HM | S | U | HM | 0.4 | 0.5 | 0.6 | | S | U | HM | S | U | HM |
| RRFS | 17.6 | 14.3 | 11.3 | 2.2 | 19.1 | 5.8 | 8.9 | 10.2 | 1.6 | 2.8 | 17.5 | 14.4 | 11.5 | 2.9 | 71.4 | 14.2 | 23.7 | 47.1 | 2.2 | 4.2 |
| ContrastZSD | 29.0 | 27.1 | 25.9 | 4.1 | 27.6 | 13.9 | 18.5 | 16.8 | 2.9 | 4.9 | 28.7 | 25.4 | 23.9 | 6.0 | 69.1 | 12.2 | 20.7 | 41.6 | 2.8 | 5.2 |
| DescReg | **45.9** | **43.0** | **40.1** | **8.3** | **28.0** | **12.8** | **17.6** | **17.1** | **5.8** | **8.7** | **37.3** | **34.4** | **29.6** | **8.5** | **83.8** | **29.9** | **44.0** | **68.7** | **4.7** | **8.8** |

Table 2: Performance of our proposed model on xView and DOTA datasets for ZSD and GZSD settings.

| Setting | Method | DIOR | | | | DOTA | | | | xView | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | airport | bask. f. | gr. tra. f. | windmill | tenn. c. | heli. | soccer. field | swim. pool | heli. | bus | pic. track | tru. tra. w/ tox tra. | mar. vessel | motorb. | barge | reach stacker | mobile crane | scraper | excavator | ship. cont. |
| GZSD | RRFS | 3.1 | 2.0 | 6.3 | 0.0 | 4.4 | 0.0 | 4.5 | 0.0 | 0.8 | 2.0 | 0.0 | 0.0 | 3.9 | 5.5 | 5.2 | **0.1** | 0.0 | **1.1** | 0.0 | **0.4** |
| | ContrastZSD | **5.2** | 2.1 | 8.1 | 0.0 | 3.5 | **2.9** | 4.8 | 0.0 | 8.1 | **5.5** | **0.1** | 1.2 | 6.3 | **9.7** | 1.2 | 0.0 | 0.0 | 0.1 | 3.1 | 0.0 |
| | DescReg | 0.0 | **9.2** | **20.0** | **2.4** | **9.1** | 0.1 | **9.5** | 0.0 | **21.9** | 4.5 | 0.0 | **6.1** | **13.3** | 9.1 | **8.2** | 0.0 | 0.0 | 0.4 | **6.1** | 0.0 |
| ZSD | RRFS | **12.3** | 6.2 | 19.7 | 0.6 | 5.4 | 0.1 | 6.1 | 0.0 | 0.1 | 2.4 | 0.0 | 0.1 | 6.9 | 1.5 | 9.2 | **0.5** | 0.0 | **5.1** | 0.0 | 0.0 |
| | ContrastZSD | 9.7 | 3.9 | 21.2 | 0.1 | 7.4 | **4.5** | 11.9 | 0.0 | 14.1 | **5.7** | 0.0 | 2.8 | 7.3 | **9.7** | 1.1 | 0.1 | 0.0 | 0.6 | 7.6 | 0.0 |
| | DescReg | 0.1 | **10.9** | **45.7** | **3.9** | **11.3** | 0.4 | **22.2** | **0.1** | **36.0** | 4.9 | 0.1 | **7.5** | **19.8** | 9.1 | **10.2** | 0.0 | **0.4** | 0.9 | **10.3** | 0.0 |

Table 3: Class-wise AP comparison of different methods on unseen classes of three aerial image datasets.

proposed DescReg can be integrated into the framework for generating more discriminative samples. As shown in Tab. 4, by augmenting with DescReg, the best-reporting generative method of RRFS is improved on PASCAL VOC dataset. Specifically, with DescReg, the mAP performance for unseen classes is improved from 65.5% to 66.4% on ZSD setting, and from 49.1% to 50.4% on GZSD setting.

| Method | ZSD | GZSD | | |
|---|---|---|---|---|
| | | S | U | HM |
| SAN (2018) | 59.1 | 48.0 | 37.0 | 41.8 |
| HRE (2018) | 54.2 | **62.4** | 25.5 | 36.2 |
| BLC (2020) | 55.2 | 58.2 | 22.9 | 32.9 |
| RRFS (2022) | 65.5 | 47.1 | 49.1 | 48.1 |
| RRFS w/. DescReg | **66.4** | 47.1 | **50.4** | **48.6** |

Table 4: ZSD and GZSD performance of the generative method on the PASCAL VOC dataset.

**DescReg with Other Detection Architectures** In addition to Faster R-CNN, we further validate DescReg on the one-stage YOLOv8 (Jocher, Chaurasia, and Qiu 2023) and the multi-stage Cascaded R-CNN (Cai and Vasconcelos 2018). The results are shown in Tab. 5. Our method applies well to the two detection models, e.g. with 15.6% ZSD mAP and Cascased R-CNN and 6.4% ZSD mAP on YOLOv8 which achieves 64 FPS inference speed.

## Analysis

We conduct several ablation studies and experimental analyses to better understand how the proposed method works.

| Architecture | ZSD | GZSD | | | FPS |
|---|---|---|---|---|---|
| | | S | U | HM | |
| Faster-RCNN | 15.2 | 68.7 | 7.9 | 14.2 | 11 |
| Cascaded-RCNN | **15.6** | **70.0** | **8.1** | **14.5** | 8 |
| YOLOv8-s | 6.4 | 49.9 | 4.2 | 7.7 | **64** |

Table 5: Performance with different detection architectures on the DIOR dataset. FPS denotes frame per seconds.

Please refer to supplementary for qualitative results.

**Ablation Study on the Proposed Triplet Loss** As shown in Tab. 6, when replacing the semantic class embeddings with the visual description embeddings, the baseline performance is improved but the improvement is limited (e.g., 1.0% on ZSD mAP). This result means naively incorporating the visual description information as the class semantic representation cannot help much due to the representation gap between semantic feature space and visual feature space. Additionally, by applying the proposed inter-class triplet loss, the performance is significantly improved (from 7.1% to 10.1% on ZSD mAP) which indicates that simple similarity-based triplet regularization could improve the zero-shot detection performance. By further introducing the proposed similarity-aware margin, the ZSD mAP is improved by 5.0% and HM is improved by 5.6%, meaning the adaptive margin helps better regularize the class representation space. We also observe the temperature value of 0.03 achieves the best performance, which is slightly higher than 0.01 and 0.05. Based on the best-performing model, further adding the visual description embeddings cannot offer im-

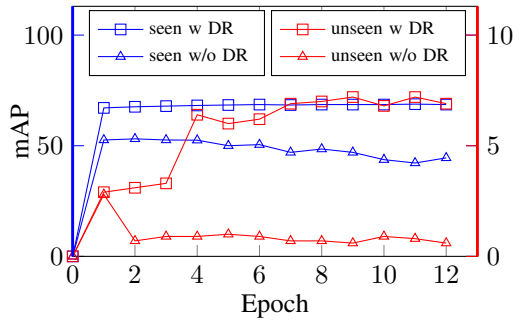


Figure 3: Learning dynamics of DescReg. w/wo DescReg denotes DescReg and baseline without DescReg.

provement, indicating our method may already incorporate the visual characteristics into the embeddings through structural similarity regularization. Fig. 3 shows how the learning dynamics of seen and unseen classes, apparently, with DescReg, the performance on both seen and unseen classes are higher and the learning process is more stable.

| S→V | Desc-Softmax | Desc-Adaptive Margin | ZSD | HM |
|---|---|---|---|---|
| | | | 6.1 | 5.3 |
| ✓ | | | 7.1 | 5.9 |
| | ✓(0.03) | | 10.1 | 8.2 |
| | ✓(0.01) | ✓ | 15.1 | 13.8 |
| | ✓(0.03) | ✓ | 15.2 | **14.2** |
| | ✓(0.05) | ✓ | 14.5 | 13.6 |
| ✓ | ✓(0.03) | ✓ | **15.3** | 13.9 |

Table 6: Ablation study of the proposed similarity-aware triplet loss. S→V means replacing the semantic embeddings with visual description embeddings. Desc-Softmax and Desc-Adaptive-Margin denote the proposed self-excluding Softmax and the similarity-aware triplet loss. The numbers in the parentheses are the temperature used in the Softmax.

**Effect of Varying Descriptions** We investigate how sensitive is DescReg to the input visual descriptions by varying the description sources. We evaluate how different human and GPT-4 (OpenAI 2023) description inputs affect the zero-shot detection performance. As shown in Tab. 7, when focusing on the semantics, the performance of both human and GPT-4 descriptions is low (e.g., 5.0% ZSD mAP for human input and 6.9% mAP for GPT-4 input). The reason is that simple semantic description contains much fewer visual details of the objects. When switching to descriptions that focus on visual details from an aerial view, the performance is significantly improved by more than 8.0% in ZSD mAP and 7.0% in HM, benefiting from the visual details that generate effective similarity measures. We also test how sensitive the method works with different descriptions of varying lengths. The result shows that our method is not very sensitive to the description length. In addition, we observe that GPT-generated descriptions offer higher performance than that of human inputs. While we did not dedicatedly optimize the human input, the result shows the application of large language models in ZSD is very efficient.

| - | Method | ZSD | HM |
|---|---|---|---|
| **Human** | semantic | 5.0 | 4.7 |
| | aerial | 13.1 | 11.9 |
| **GPT-4** | semantic | 6.9 | 6.1 |
| | aerial-long | **15.2** | **14.2** |
| | aerial-medium | 15.1 | 13.9 |
| | aerial-short | 13.5 | 13.2 |

Table 7: Varying descriptions. We acquire visual descriptions through both Human and GPT-4. *semantic* means simply describing the object class while *aerial* means focusing on the visual appearance in aerial images. *long*, *medium*, and *short* denote descriptions with varying lengths.

## Conclusion

In this paper, we investigate the zero-shot object detection (ZSD) problem in the context of aerial images. We identified the weak semantic-visual correlation problem of aerial objects and propose to learn stronger visually-aligned class representations with external visual descriptions in text format. Our method is extensively validated on three challenging aerial object detection datasets and shows significantly improved performance to the prior ZSD methods. To the best of our knowledge, we are the first to conduct a comprehensive study on zero-shot aerial object detection. we hope our method and newly established experimental setups provide a baseline for future research.

**Limitations and Future Work** While our method significantly improves the baselines, we note the performance on unseen classes is still low. The major challenge arises from the strong inter-class confusion and background confusion among aerial objects, which is further exacerbated by the small object size. While our method mitigates these problems, two future directions could further address them: 1) The non-uniform spatial processing approaches (Recasens et al. 2018; Yang et al. 2019) could be explored to amplify the small object signal for improved zero-shot recognition. 2) Based on our proposed regularization, other label-efficient methods could be incorporated to improve the performance, e.g. few-shot approach and open-vocabulary detection approach (Kang et al. 2019; Wang 2023).

## Acknowledgments

This work is supported by the Key Program of National Science Foundation of China under Grant 61836006, the Fundamental Research Funds for the Central Universities under Grant YJ202342 and 1082204112364, the National Science Foundation of China under Grant 62106161, the Key R&D Program of Sichuan Province under Grant 2022YFN0017 and 2023YFG0278 and Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education.

## References


Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2015. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-Shot Object Detection. In *European Conference on Computer Vision*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Cheng, G.; Han, J.; Zhou, P.; and Xu, D. 2019. Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. *IEEE Transactions on Image Processing*, 28: 265–278.

Cheng, G.; Zhou, P.; and Han, J. 2016. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 54: 7405–7415.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2018. Zero-Shot Object Detection by Hybrid Region Embedding. In *British Machine Vision Conference*.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2019. Learning Visually Consistent Label Embeddings for Zero-Shot Learning. *2019 IEEE International Conference on Image Processing (ICIP)*, 3656–3660.

Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; and Qin, H. 2020. A global-local self-adaptive network for drone-view object detection. *IEEE Transactions on Image Processing*, 30: 1556–1569.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M. Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. 2021. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7778–7796.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2584–2591.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Han, J.; Ding, J.; Li, J.; and Xia, G.-S. 2021. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Hayat, N.; Hayat, M.; Rahman, S.; Khan, S. H.; Zamir, S. W.; and Khan, F. S. 2020. Synthesizing the Unseen for Zero-shot Object Detection. In *Asian Conference on Computer Vision*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, P.; Han, J.; Cheng, D.; and Zhang, D. 2022. Robust Region Feature Synthesizer for Zero-Shot Object Detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7612–7621.

Jasani, B.; and Mazagonwalla, A. 2019. Skeleton based Zero Shot Action Recognition in Joint Pose-Language Semantic Space. *ArXiv*, abs/1911.11344.

Jocher, G.; Chaurasia, A.; and Qiu, J. 2023. YOLO by Ultralytics.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8420–8429.

Khandelwal, S.; Nambirajan, A.; Siddiquie, B.; Eledath, J.; and Sigal, L. 2023. Frustratingly Simple but Effective Zero-shot Detection and Segmentation: Analysis and a Strong Baseline. *arXiv preprint arXiv:2302.07319*.

Koyun, O. C.; Keser, R. K.; Akkaya, I. B.; and Töreyin, B. U. 2022. Focus-and-Detect: A small object detection framework for aerial images. *Signal Processing: Image Communication*, 104: 116675.

Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M. K.; Bulatov, Y.; and McCord, B. 2018. xView: Objects in Context in Overhead Imagery. *ArXiv*, abs/1802.07856.

Li, C.; Yang, T.; Zhu, S.; Chen, C.; and Guan, S. 2020. Density map guided object detection in aerial images. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 190–191.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2019a. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ArXiv*, abs/1909.00133.

Li, W.; Chen, Y.; Hu, K.; and Zhu, J. 2022. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1829–1838.

Li, Z.; Yao, L.; Zhang, X.; Wang, X.; Kanhere, S. S.; and Zhang, H. 2019b. Zero-Shot Object Detection with Textual Descriptions. In *AAAI Conference on Artificial Intelligence*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, X.; Sun, X.; Diao, W.; Mao, Y.; Li, J.; Zhang, Y.; Wang, P.; and Fu, K. 2023. Few-shot object detection in aerial imagery guided by text-modal knowledge. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–19.

Meethal, A.; Granger, E.; and Pedersoli, M. 2023. Cascaded Zoom-in Detector for High Resolution Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2045–2054.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mishra, A.; Verma, V. K.; Reddy, M. S. K.; Subramaniam, A.; Rai, P.; and Mittal, A. 2018. A Generative Approach to Zero-Shot and Few-Shot Action Recognition. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 372–380.

OpenAI. 2023. GPT-4. https://chat.openai.com/. Accessed: 2024-01-24.

Paz-Argaman, T.; Atzmon, Y.; Chechik, G.; and Tsarfaty, R. 2020. Zest: Zero-shot learning from text descriptions using textual similarity and visual summarization. *arXiv preprint arXiv:2010.03276*.

Rahman, S.; Khan, S.; and Barnes, N. 2020. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11932–11939.

Rahman, S.; Khan, S.; and Porikli, F. 2018. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, 547–563. Springer.

Recasens, A.; Kellnhofer, P.; Stent, S.; Matusik, W.; and Torralba, A. 2018. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 51–66.

Redmon, J.; and Farhadi, A. 2017. YOLO9000: better, faster, stronger. *arXiv preprint*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Sarma, S.; Kumar, S.; and Sur, A. 2022. Resolving Semantic Confusions for Improved Zero-Shot Detection. *arXiv preprint arXiv:2212.06097*.

Tang, C.; He, Z.; Li, Y.; and Lv, J. 2021. Zero-shot learning via structure-aligned generative adversarial network. *IEEE transactions on neural networks and learning systems*, 33(11): 6749–6762.

Tang, C.; Lv, J.; Chen, Y.; and Guo, J. 2019. An angle-based method for measuring the semantic similarity between visual and textual features. *Soft Computing*, 23: 4041–4050.

Tang, C.; Yang, X.; Lv, J.; and He, Z. 2020. Zero-shot learning by mutual information estimation and maximization. *Knowledge-Based Systems*, 194: 105490.

Wang, Q.; and Chen, K. 2017. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 124: 356–383.

Wang, T. 2023. Learning to detect and segment for open vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7051–7060.

Wolf, S.; Meier, J.; Sommer, L.; and Beyerer, J. 2021. Double head predictor based few-shot object detection for aerial imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 721–731.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S. J.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2017. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3974–3983.

Yan, C.; Chang, X.; Luo, M.; Liu, H.; Zhang, X.; and Zheng, Q. 2022. Semantics-guided contrastive network for zero-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, C.; Huang, Z.; and Wang, N. 2022. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *IEEE Conference on computer vision and pattern recognition*, 13668–13677.

Yang, F.; Fan, H.; Chu, P.; Blasch, E.; and Ling, H. 2019. Clustered object detection in aerial images. In *IEEE International conference on computer vision*, 8311–8320.

Zhang, L.; Wang, X.; Yao, L.; Wu, L.; and Zheng, F. 2020a. Zero-shot object detection via learning an embedding from semantic space to visual space. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Zhang, Z.; Jiang, R.; Mei, S.; Zhang, S.; and Zhang, Y. 2020b. Rotation-Invariant Feature Learning for Object Detection in VHR Optical Remote Sensing Images by Double-Net. *IEEE Access*, 8: 20818–20827.

Zhao, S.; Gao, C.; Shao, Y.; Li, L.; Yu, C.; Ji, Z.; and Sang, N. 2020. GTNet: Generative Transfer Network for Zero-Shot Object Detection. *ArXiv*, abs/2001.06812.

Zheng, Y.; Huang, R.; Han, C.; Huang, X.; and Cui, L. 2020. Background Learnable Cascade for Zero-Shot Detection.

Zhu, P.; Wang, H.; and Saligrama, V. 2019. Don't Even Look Once: Synthesizing Features for Zero-Shot Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11690–11699.

Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; and Ling, H. 2021. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7380–7399.