SAUI: Scale-Aware Unseen Imagineer for Zero-Shot Object Detection

Jiahao Wang[1], Caixia Yan[1*], Weizhan Zhang[1*], Huan Liu[1], Hao Sun[2], Qinghua Zheng[1]

[1]School of Computer Science and Technology, MOEKLINNS Laboratory, Xi'an Jiaotong University
[2]China Telecom Artificial Intelligence Technology Co.Ltd
uguisu@stu.xjtu.edu.cn, {yancaixia, zhangwzh, huanliu, qhzheng}@xjtu.edu.cn,
sunh10@chinatelecom.cn

Abstract

Zero-shot object detection (ZSD) aims to localize and classify unseen objects without access to their training annotations. As a prevailing solution to ZSD, generation-based methods synthesize unseen visual features by taking seen features as reference and class semantic embeddings as guideline. Although previous works continuously improve the synthesis quality, they fail to consider the scale-varying nature of unseen objects. The generation process is preformed over a single scale of object features and thus lacks scale-diversity among synthesized features. In this paper, we reveal the scale-varying challenge in ZSD and propose a **S**cale-**A**ware **U**nseen **I**magineer (**SAUI**) to lead the way of a novel scale-aware ZSD paradigm. To obtain multi-scale features of seen-class objects, we design a specialized coarse-to-fine extractor to capture features through multiple scale-views. To generate unseen features scale by scale, we innovate a Series-GAN synthesizer along with three scale-aware contrastive components to imagine separable, diverse and robust scale-wise unseen features. Extensive experiments on PASCAL VOC, COCO and DIOR datasets demonstrate SAUI's better performance in different scenarios, especially for scale-varying and small objects. Notably, SAUI achieves the new state-of-the-art performance on COCO and DIOR.

Introduction

Zero-Shot Object Detection (ZSD) (Bansal et al. 2018; Rahman, Khan, and Porikli 2018) is the task that aims to detect novel objects without any access to their corresponding annotations. Besides seen-class objects involved in the labeled training dataset, ZSD models can also detect unseen-class objects beyond the dataset. Thus, it avoids the laborious pre-task of annotating and simultaneously contributes to the detection of rarely captured objects, such as endangered animals and novelty items. Generally, the unseen-class object scales in ZSD vary intensively due to different poses, shapes or shooting angles. Since models are not allowed to learn the real scale-diverse distribution of unseen features, detecting scale-varying unseen-class objects still remains a key challenge in ZSD.

In recent years, ZSD has achieved tremendous advancements (Zhao et al. 2020; Hayat et al. 2020; Zheng et al. 2020; Yan et al. 2022; Huang et al. 2022), which is indeed dominated by mapping-based and generation-based approaches. Specifically, mapping-based methods (Zheng et al. 2020; Yan et al. 2022) manage to align the visual features to class embeddings and perform the nearest neighbor search to find their classes. However, lack of unseen visual features causes severe bias towards seen objects and thus leads to poor performance or even mode collapse when classifying (Khan et al. 2019). For this issue, generation-based methods (Zhao et al. 2020; Hayat et al. 2020; Huang et al. 2022) are proposed to synthesize unseen visual features based on the corresponding unseen class embeddings. Both real seen features and synthesized unseen features can contribute to the training of classifier and thereby significantly alleviate the bias problem towards seen classes.

Existing generation-based ZSD methods have utilized Feature Pyramid Network (FPN) (Lin et al. 2017) to process seen objects of different scales and then leverage their

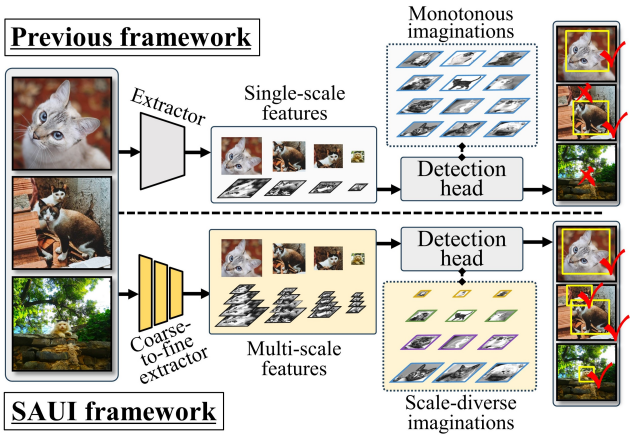


Figure 1: Comparison between previous framework and the proposed SAUI framework, where different colored boxes denote different scale-levels. Top: Previous generative methods for ZSD extract single-scale features and then imagine, which neglects the diversity of different scale-levels; Bottom: Our SAUI extracts coarse-to-fine features through multiple scale-views and imagine diverse scale-wise features to improve scale-varying object detection.

*Corresponding authors

features to train the classifier. Thus, they do well in the detection of scale-varying seen objects. When it comes to the unseen classes, we certainly hope the generator can imagine scale-wise diverse features of unseen objects. But unfortunately, the generators in current methods neglect the concept of object scale. In previous works, the extractor processes each region proposal and obtains each seen feature through one specific scale-view, as shown in Fig. 1. Though two seen features may come from two different scale-views, the generator treats them as the same and then imagines unseen features regardless of the scale information. This leads to lack of scale-diversity of the synthesized features, and thus further diminishes the ZSD performance.

When human see things, we usually match the scale-specific appearance of visual objects with our past memories. The scales won't confuse us because our brain memorizes what things look like in multiple scale-views. We focus on the scale-invariant commonality and scale-wise individuality (Han et al. 2017; Nercessian, Panetta, and Agaian 2013) and thus can imagine things in other possible views. Inspired by this, we propose Scale-Aware Unseen Imagineer (SAUI), a novel scale-aware generation-based ZSD paradigm, which sees and imagines visual objects guided by the concept of scale. As shown in Fig. 1, our generator synthesizes scale-diverse unseen features to fairly detect scale-varying unseen objects. To achieve this goal, we specifically design a coarse-to-fine extractor and a scale-aware synthesizer for SAUI. The extractor captures visual features through multiple scale-view channels, where each scale-view retains information of one scale-level. For the synthesizer, inspired by the pyramid-like network, we pile up the generators and discriminators to form a Series-GAN. Then, we formulate the coarse-to-fine synthesis as a sequence-to-sequence problem to generate features scale by scale. To stabilize and optimize the synthesis process, we further develop three multi-scale contrastive components to simultaneously diverge and separate intra-scale features as well as co-relate inter-scale features. Extensive experiments demonstrate the superiority of our scale-aware method in both ZSD and GZSD settings.

The main contributions of this paper are as follows:

- We reveal the scale-varying challenge in current ZSD approaches and first introduce the scale-awareness to facilitate generation-based ZSD methods.
- We build up a novel scale-aware generation paradigm for ZSD, which is equipped with a coarse-to-fine extractor to capture scale-aware details and a Series-GAN synthesizer to generate scale-diverse imaginations.
- We formulate the training of the Series-GAN synthesizer as a sequence-to-sequence problem, and propose three contrastive components to make the synthesized multi-scale features separable, diverse and robust.
- Extensive and comprehensive experiments on PASCAL VOC, COCO and DIOR datasets demonstrate the superior detection performance of SAUI, especially for small objects and scale-varying objects.

Related Work

Zero-Shot Object Detection. ZSD, first introduced in (Rahman, Khan, and Porikli 2018; Bansal et al. 2018), aims to perform detection of unseen-class objects. Different from Open World Detection(Joseph et al. 2021) and Open Vocabulary Detection(Gu et al. 2021), ZSD classifies unseen objects to their corresponding classes without any human annotations or extra training data. To accomplish this task, mapping-based methods optimize the alignment between the visual space and the pre-defined semantic space (Li et al. 2019; Rahman et al. 2020; Yan et al. 2022). More recently, generation-based methods are proposed to alleviate the severe bias issue of mapping-based methods. GTNet (Zhao et al. 2020) first presents a synthesizer-assisted paradigm to balance the classifier training with synthesized unseen-class features. SU (Hayat et al. 2020) improves the framework by simplifying the synthesizer and instead use loss components to manipulate the synthesis. RRFS (Huang et al. 2022) adds more components to reinforce the intra-class diversity and inter-class separability. However, to the best of our knowledge, previous works hardly ever take object scale information into account. This gives rise to our SAUI, a scale-aware zero-shot object detection framework.

Scale-Aware Object Detection. Object detection has achieved tremendous advancements (Girshick 2015; Ren et al. 2017; Dosovitskiy et al. 2021; Carion et al. 2020) in recent years and varying object scale remains a main challenge. Many efforts (Lin et al. 2017; Zhu et al. 2021; Liu et al. 2021) have laid a solid foundation to address the scale challenge. In general, they utilize multi-scale feature representations to empower the detectors with scale-awareness. FPN (Lin et al. 2017) constructs a top-down architecture with lateral connections to capture features through multiple scale-views. Deformable DETR (Zhu et al. 2021) proposes the deformable attention module to aggregate multi-scale features, especially improving the detection performance of small objects. Swin Transformer (Liu et al. 2021) presents a hierarchical representation which provides scale adaptability for ViT (Dosovitskiy et al. 2021). Driven by their success, we introduce scale-awareness into generation-based ZSD methods to generate scale-diverse features.

The Proposed Method

Problem Definition

Let's start with the definition of ZSD task. Given seen-class set $\mathcal{S}$ of C_s different class names and unseen-class set $\mathcal{U}$ of C_u class names, $\mathcal{S} \cap \mathcal{U} = \varnothing$. During training, the training set $\mathcal{X}^s = \{\mathcal{I}_i^s\}_{i=1}^{N_s}$ consists of N_s annotated images, where each image contains several instances of seen-class objects. Each object is annotated with bounding boxes $\mathbf{b} \in \mathbb{R}^4$ and its label $y \in \mathcal{S}$. Along with the image set, m-dimensional semantic embeddings of each class are provided and denoted as $\{\mathbf{w}_i^s\}_{i=1}^{C_s}$ for seen classes and $\{\mathbf{w}_i^u\}_{i=1}^{C_u}$ for unseen classes. During inference, the test set is denoted as $\mathcal{X}^u = \{\mathcal{I}_i^u\}_{i=1}^{N_u}$. For typical ZSD task, we have to detect objects of the unseen classes in the test set. While for the GZSD task, detection of both seen and unseen classes is expected.

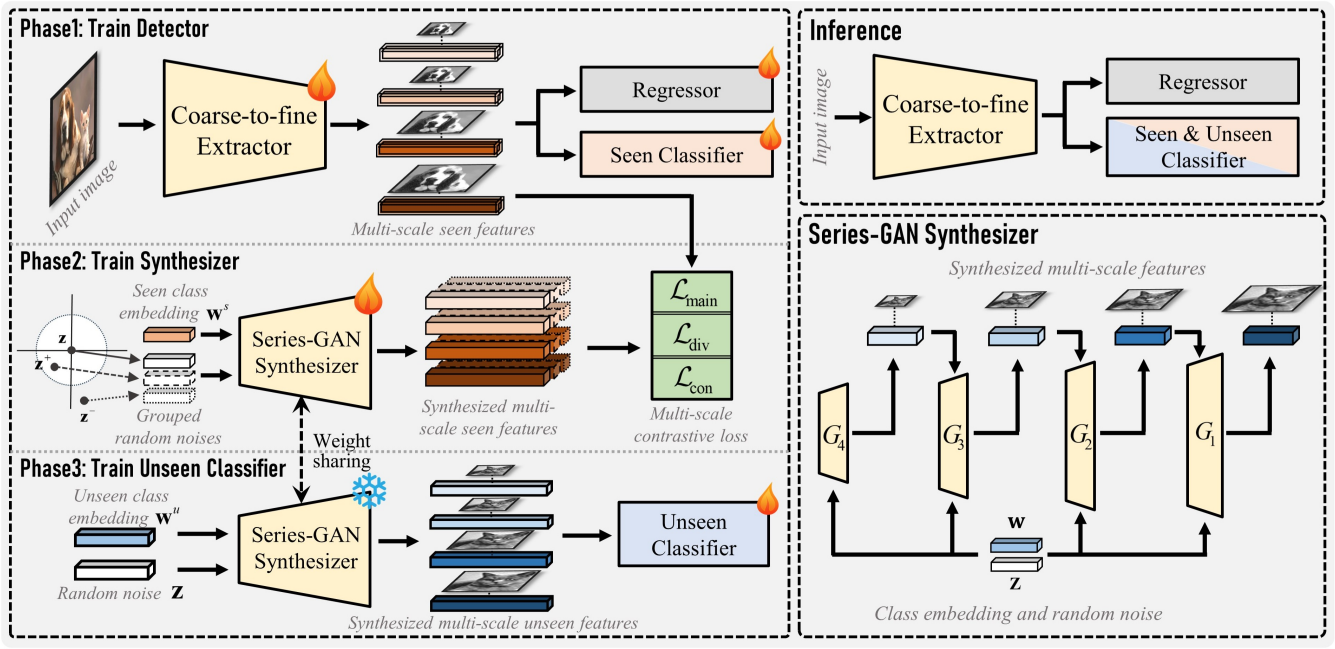


Figure 2: The overall architecture of the proposed SAUI framework. The training includes train detector with seen annotations, train synthesizer with seen features and train unseen classifier with synthesized unseen features. The core models of our method are the coarse-to-fine extractor, the series-GAN synthesizer and multi-scale contrastive loss components. Tuning and freezing weights are denoted by fire and snowflake marks. Discriminators are not shown for simplification.

Scale-Aware Unseen Imagineer

Overview. Fig. 2 illustrates the framework of SAUI. Our method captures the coarse-to-fine visual features of seen objects and then sketches unseen features by leveraging multi-scale information from seen objects. We employ Faster-RCNN (Girshick 2015) as the basic detection model with ResNet-101 (He et al. 2016) backbone. The training procedure can be summarized as 3 phases: 1) Train the detector on seen image set. 2) Extract multi-scale region features of seen-class objects using the trained coarse-to-fine extractor; then train the Series-GAN synthesizer with multi-scale contrastive components to learn the alignment between the multi-scale seen features and their corresponding word embeddings. 3) Synthesize multi-scale features of unseen-class objects; then train a classifier specified for unseen objects. After training, we can update the classifier of the origin detector to form a fresh ZSD detector for inference.

In general, our novelty lies in the design of coarse-to-fine feature extractor, the structure of Series-GAN and the proposal of three multi-scale contrastive components.

Coarse-to-fine Feature Extractor. To retain as many scale features as possible, we design a new coarse-to-fine detector structure. Given a single input image $\mathcal{I}_i^s$, vanilla Faster-RCNN extracts features with backbone and FPN neck, generates N_p region proposals with RPN and obtains features of proposals $\{\mathbf{f}_i\}_{i=1}^{N_p}$ with RoI pooling for further recognition. Notably, although FPN extracts image features through multiple scale-views like a pyramid, every $\mathbf{f}_i$ comes from one specific scale-view chosen by rules. In contrast, we add a series of scale-view channels to form a coarse-to-fine feature extractor. Each channel corresponds to a RoI Pooling model connected to one scale-level of the FPN pyramid that extracts single-scale region features. All channels are aggregated to jointly compose multi-scale features $\{\mathbf{f}_{i,j}\}_{i=1,j=1}^{N_p,N_v}$, where N_v refers to the number of scale-view channels. To distinguish seen and unseen instances, we denote $\mathbf{f}_{i,j}^s$ and $\mathbf{f}_{i,j}^u$ as seen and unseen features, respectively. Moreover, features with smaller j come from lower level of the pyramid and tend to be finer.

Series-GAN Synthesizer. With coarse-to-fine seen features $\{\mathbf{f}_{i,j}^s\}_{i=1,j=1}^{N_p,N_v}$, vanilla GAN synthesizer is not suitable anymore. To leverage the inherent co-relations of different scale-views, we construct a series structure of GAN. Since the features no longer preserve spatial relations as the feature maps do, we cannot use the Laplacian Pyramid (Denton et al. 2015) directly as our synthesizer. Thus, we rethink its essence and apply the top-to-bottom generation idea to our work. Specially, we construct a series of single-scale generators $\{\mathbf{G}_j\}_{j=1}^{N_v}$ and discriminators $\{\mathbf{D}_j\}_{j=1}^{N_v}$ to form a Series-GAN Synthesizer. The operating process of the synthesizer contains two stages: sampling and training. During sampling, the coarsest $\mathbf{G}_{N_v}$ takes word embedding $\mathbf{w}$ and a n-dimensional random noise vector $\mathbf{z}_{N_v}$ as inputs and synthesize coarsest feature $\tilde{\mathbf{f}}_{N_v}$. The following $\mathbf{G}_j$ takes in $\mathbf{w}$, $\mathbf{z}_j$ and the former feature $\tilde{\mathbf{f}}_{j+1}$ to synthesize a finer feature $\tilde{\mathbf{f}}_j$, and so on, until the finest $\tilde{\mathbf{f}}_1$ is synthesized. During training, we apply sequence-to-sequence strategy and the process will be detailed below.

Sequence-to-sequence Training. The Series-GAN sees the coarse visual features, ponders the semantic embedding of target classes, and then sketches finer details on the vague imaginations. Naturally, if we regard the features as sequences and the coarse-to-fine relation as the former-to-latter order, this problem can be roughly transferred to a sequence-to-sequence problem. Thus, we can apply sequence-to-sequence learning strategy to optimize its training and inference.

During training, the generator $\mathbf{G}_j$ always takes in the real seen feature $\mathbf{f}_j^s$ rather than the synthesized feature $\tilde{\mathbf{f}}_j^s$, while every discriminator $\mathbf{D}_j$ takes in $\mathbf{z}_j$ and tries to separate the corresponding $\tilde{\mathbf{f}}_j^s$ apart from real features $\mathbf{f}_j^s$. For one pair of a generator and a discriminator, the learning objective function is:

$$\min_{\mathbf{G}_j}\max_{\mathbf{D}_j} \mathcal{L}_{\text{WGAN}} + \alpha_1\mathcal{L}_{\text{cls}} + \alpha_2\mathcal{L}_{\text{main}} + \alpha_3\mathcal{L}_{\text{div}} + \alpha_4\mathcal{L}_{\text{con}}, \tag{1}$$

where α_1, α_2, α_3 and α_4 are hyper-parameters weighting different components to cooperate. Specifically, we use Wasserstein GAN loss $\mathcal{L}_{\text{WGAN}}$ (Arjovsky, Chintala, and Bottou 2017) to align the synthesized feature distributions to the real real feature distributions. Given synthesized seen features $\tilde{\mathbf{f}}_j^s = \mathbf{G}(\mathbf{w}_j^s, \mathbf{z}_j)$, the Wasserstein loss can be formulated as:

$$\mathcal{L}_{\text{WGAN}} = \mathbb{E}[\mathbf{D}_j(\mathbf{f}_j^s, \mathbf{w}_j^s)] - \mathbb{E}[\mathbf{D}_j(\tilde{\mathbf{f}}_j^s, \mathbf{w}_j^s)] - \lambda_{\text{w}}\mathbb{E}[(\left\|\nabla_{\hat{\mathbf{f}}_j^s}\mathbf{D}_j(\hat{\mathbf{f}}_j^s, \mathbf{w}_j^s)\right\|_2 - 1)^2], \tag{2}$$

where $\hat{\mathbf{f}}_j^s = \beta\mathbf{f}_j^s + (1-\beta)\tilde{\mathbf{f}}_j^s$ is a momentum term for regulating the gradients with $\beta \sim \mathcal{N}(0,1)$. The classification loss $\mathcal{L}_{\text{cls}}$, which preserves the mapping relations between seen visual space and seen semantic space, is described as:

$$\mathcal{L}_{\text{cls}} = -\mathbb{E}[\log p(y|\mathbf{G}(\mathbf{w}_j^s, \mathbf{z}_j); \phi_{\text{cls}})]. \tag{3}$$

For a further explanation, the classifier from the trained Faster-RCNN, *i.e.*, ϕ_{cls}, is frozen throughout the training process of the synthesizer.

Multi-scale Contrastive Imagination

In a well-trained visual feature space for seen objects, feature vectors of instances from the same class tend to cluster while those from different classes tend to disperse, which shows their separability. Besides, in one specific class, feature vectors from different instances show slight discrepancy due to their diversity. We term the two observations above as *outline* and *detail*. In this work, since the visual feature space consists of a set of features from different scale-views, we should also pay attention to the co-construction between features of different scale-levels.

In order to precisely construct the cross-scale visual feature space generated by the synthesizer, as shown in Fig. 3, we introduce three multi-scale contrastive components, including Intra-Scale Outline Maintaining, Intra-scale Detail Diverging and Inter-scale Mutual Constructing. They simultaneously repel and attract different synthesized features from different classes and scale-views towards the positions

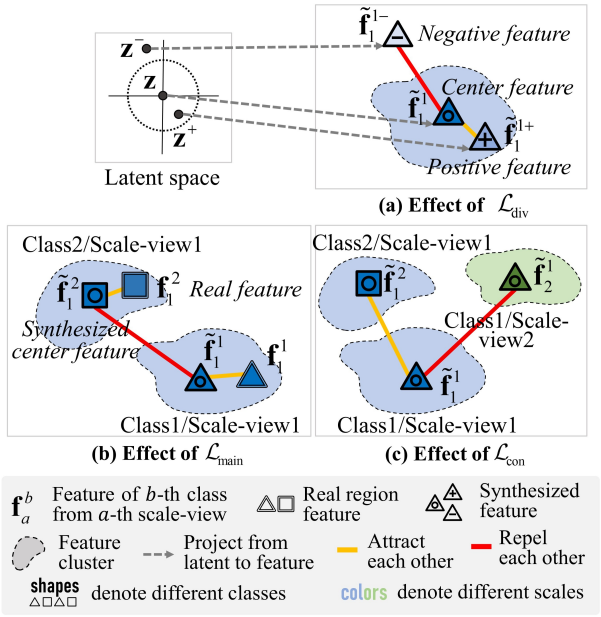


Figure 3: Illustration of the effect of (a) $\mathcal{L}_{\text{div}}$, (b) $\mathcal{L}_{\text{main}}$ and (c) $\mathcal{L}_{\text{con}}$. In the feature space, $\mathcal{L}_{\text{div}}$ encourages synthesized features from the same class and scale-view to preserve the inner diversity. $\mathcal{L}_{\text{main}}$ encourages synthesized features to align to real ones and clusters features from the same class. $\mathcal{L}_{\text{con}}$ clusters features from the same scale-view.

they should naturally be. Notice that $\mathbf{f}_j$ below represents features of seen classes if not specified.

Intra-scale Outline Maintaining. To maintain the separability of different classes in each scale-view, we introduce Intra-Scale Outline Maintaining loss over the synthesized feature vectors, *i.e.*,

$$\mathcal{L}_{\text{main}} = -\mathbb{E}[\log \frac{\exp(\tilde{\mathbf{f}}_j \cdot \mathbf{g}_j^+/\tau)}{\exp(\tilde{\mathbf{f}}_j \cdot \mathbf{g}_j^+/\tau) + \sum_{n=1}^{N}\exp(\tilde{\mathbf{f}}_j \cdot \mathbf{g}_{j,n}^-/\tau)}], \tag{4}$$

where $\mathbf{g}$ is sampled from a visual feature pool consisting of synthesized features set $\{\tilde{\mathbf{f}}_j\}$ and real region features set of proposals $\{\mathbf{f}_j\}$; τ is the temperature coefficient. If we denote $y(\cdot)$ as a function that indicates the class of one feature vector, then $y(\mathbf{g}_j^+) = y(\tilde{\mathbf{f}}_j)$ while $y(\mathbf{g}_j^-) \neq y(\tilde{\mathbf{f}}_j)$. As a result, $\mathcal{L}_{\text{main}}$ forces features of different classes to repel and those from the same class to attract.

Intra-scale Detail Diverging. To diverge the peculiarities of synthesized feature vectors of the same class in each scale-view, we introduce Intra-scale Detail Diverge loss:

$$\mathcal{L}_{\text{div}} = -\mathbb{E}[\log \frac{\exp(\tilde{\mathbf{f}}_j \cdot \tilde{\mathbf{f}}_j^+/\tau)}{\exp(\tilde{\mathbf{f}}_j \cdot \tilde{\mathbf{f}}_j^+/\tau) + \sum_{n=1}^{N}\exp(\tilde{\mathbf{f}}_j \cdot \tilde{\mathbf{f}}_{j,n}^-/\tau)}], \tag{5}$$

which forces features from the same class to differ from each other based on $\tilde{\mathbf{f}}_j = \mathbf{G}(\mathbf{w}_j, \mathbf{z}_j)$, together with the

Method	Split	ZSD				GZSD					
		Recall@100			mAP	Recall@100			mAP		
		IoU=0.4	IoU=0.5	IoU=0.6	IoU=0.5	S	U	HM	S	U	HM
PL (Rahman et al. 2020)	48/17	-	43.5	-	10.1	38.2	26.3	31.2	35.9	4.1	7.4
BLC (Zheng et al. 2020)	48/17	51.3	48.8	45.0	10.6	57.6	46.4	51.4	42.1	4.5	8.2
ConZSD (Yan et al. 2022)	48/17	56.1	52.4	47.2	12.5	65.7	52.4	58.3	45.1	6.3	11.1
RRFS* (Huang et al. 2022)	48/17	58.1	53.5	47.9	13.4	59.7	58.8	59.2	42.3	13.4	20.4
TCB (Li et al. 2023)	48/17	55.5	52.4	48.1	11.4	**72.0**	52.4	60.7	**47.3**	4.9	8.9
SAUI (Ours)	48/17	**66.5**	**66.0**	**63.9**	**13.7**	64.7	**59.3**	**61.9**	43.3	**13.6**	**20.7**
PL (Rahman et al. 2020)	65/15	-	37.7	-	12.4	36.4	37.2	36.8	34.1	12.4	18.2
SU* (Hayat et al. 2020)	65/15	54.4	54.0	47.0	19.0	57.7	53.9	55.8	36.9	19.0	25.1
BLC (Zheng et al. 2020)	65/15	57.2	54.7	51.2	14.7	56.4	51.7	53.9	36.0	13.1	19.2
RSC* (Sarma et al. 2022)	65/15	-	65.1	-	20.1	58.6	64.0	61.2	37.4	20.1	26.2
TZSDC (Liu et al. 2022)	65/15	-	56.5	-	19.6	56.8	52.7	54.7	32.5	19.2	24.2
ConZSD (Yan et al. 2022)	65/15	62.3	59.5	55.1	18.6	62.9	58.6	60.7	**40.2**	16.5	23.4
RRFS* (Huang et al. 2022)	65/15	65.3	62.3	55.9	19.8	58.6	61.8	60.2	37.4	19.8	26.0
TCB (Li et al. 2023)	65/15	62.5	59.9	55.1	13.8	69.4	59.9	64.3	39.9	13.8	20.5
SAUI (Ours)	65/15	**77.9**	**77.8**	**76.5**	**22.1**	**70.7**	**69.9**	**70.3**	36.8	**21.3**	**27.0**

Table 1: Results of state-of-the-arts on COCO dataset in ZSD and GZSD settings, including mapping and generation-based methods. The best and second-best results are marked bold and underlined. The symbol "*" denotes generation-based methods.

Method	airp	trai	park	cat	bear	suit	fris	snow	fork	sand	hotd	toil	mous	tost	hair	AP_s	AP_m	AP_l	mAP
SU	10.1	48.7	1.2	64.0	**64.1**	12.2	0.7	28.0	16.4	19.4	0.1	**18.7**	1.2	0.5	0.2	3.5†	10.5†	24.3†	19.0
RRFS	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	3.6†	11.0†	25.2†	19.8
RSC	22.9	**53.3**	0.6	**64.9**	54.3	13.2	1.2	31.2	15.7	22.6	0.0	17.5	2.7	0.7	0.2	-	-	-	20.1
SAUI	**27.8**	42.4	**6.5**	53.8	53.2	**15.4**	**31.3**	28.9	8.3	**24.3**	**12.4**	14.3	**10.2**	**1.2**	**1.1**	**9.9**	**14.9**	**28.3**	**22.1**

Table 2: Class-wise AP, multi-scale AP comparison of generation-based methods on 65/15 split COCO dataset in ZSD setting with IoU=0.5. The symbol "†" denotes the reproduced results. The best results here and in the following tables are marked bold.

positive samples $\tilde{\mathbf{f}}_j^+ = \mathbf{G}(\mathbf{w}_j, \mathbf{z}_j^+)$ and negative samples $\tilde{\mathbf{f}}_j^- = \mathbf{G}(\mathbf{w}_j, \mathbf{z}_j^-)$. A set of contrastive random variables $\{\mathbf{z}_j, \mathbf{z}_j^+, \mathbf{z}_j^-\}$ are chosen by $\mathbf{z}_j^+ = \mathbf{z}_j + \rho$, where $\rho \sim \mathcal{U}[-r, r]$ and $\mathbf{z}_j^- \in \{\mathbf{z}_j^- | \mathbf{z}_j^- \sim \mathcal{N}(0,1), |\mathbf{z}_j^- - \mathbf{z}_j| > r\}$.

Inter-scale Mutual Constructing. To co-relate features of different scales, we further introduce Inter-scale Mutual Constructing loss to maximize the mutual information between features of different scales, *i.e.*,

$$\mathcal{L}_{\text{con}} = \sum_{\substack{k=1 \\ k \neq j}}^{N_v} \mathcal{L}_{\text{con}}^{j,k} + \mathcal{L}_{\text{con}}^{k,j}, \tag{6}$$

which is an accumulation of co-effects between every pair of scales. When we focus on two different scale-views j and k, the co-effect can be described as:

$$\mathcal{L}_{\text{con}}^{j,k} = -\mathbb{E}[\log \frac{\exp(\tilde{\mathbf{f}}_j \cdot \tilde{\mathbf{f}}_k^+ / \tau)}{\exp(\tilde{\mathbf{f}}_j \cdot \tilde{\mathbf{f}}_k^+ / \tau) + \sum_{n=1}^{N} \exp(\tilde{\mathbf{f}}_j \cdot \tilde{\mathbf{f}}_{k,n}^- / \tau)}], \tag{7}$$

where $\tilde{\mathbf{f}}_k$ is sampled from feature vectors set of another scale-view and $y(\tilde{\mathbf{f}}_k^+) = y(\tilde{\mathbf{f}}_k)$, $y(\tilde{\mathbf{f}}_k^-) \neq y(\tilde{\mathbf{f}}_k)$.

Experiments

Experimental Settings

Datasets. We evaluate SAUI on two typical ZSD datasets, *i.e.* PASCAL VOC 2007+2012 (Everingham et al. 2010), MS COCO 2014 (Lin et al. 2014) and one remote sensing detection dataset, *i.e.* DIOR (Li et al. 2020). For seen/unseen split manners, we adopt 16/4 split (Demirel, Cinbis, and Ikizler-Cinbis 2018) on PASCAL VOC, 48/17&65/15 split (Bansal et al. 2018) on MS COCO, and 16/4 split (Huang et al. 2022) on DIOR.

Evaluation Protocols. For PASCAL VOC and DIOR, we report mean average precision (mAP) with IoU=0.5. For MS COCO, we report both mAP with IoU=0.5 and recall@100 (RE) with IoU=0.4/0.5/0.6. Moreover, we report seen-class score (S), unseen-class score (U) and Harmonic Mean (HM) to evaluate the GZSD performance. Notably, HM scores of mAP and recall will be denoted as mAP-HM and recall-HM. To better substantiate our motivation, we follow (Lin et al. 2014) to report AP_s/AP_m/AP_l for small, medium and large objects.

Implementation Details. We set N_v to 4 so that features from 4 scale-view channels are considered. Both the generator $\mathbf{G}_j$ and discriminator $\mathbf{D}_j$ are implemented as two-layer fully-connected networks with 4096 hidden units per layer. We use CLIP text encoder (Radford et al. 2021) to obtain the semantic embeddings. Note that we never

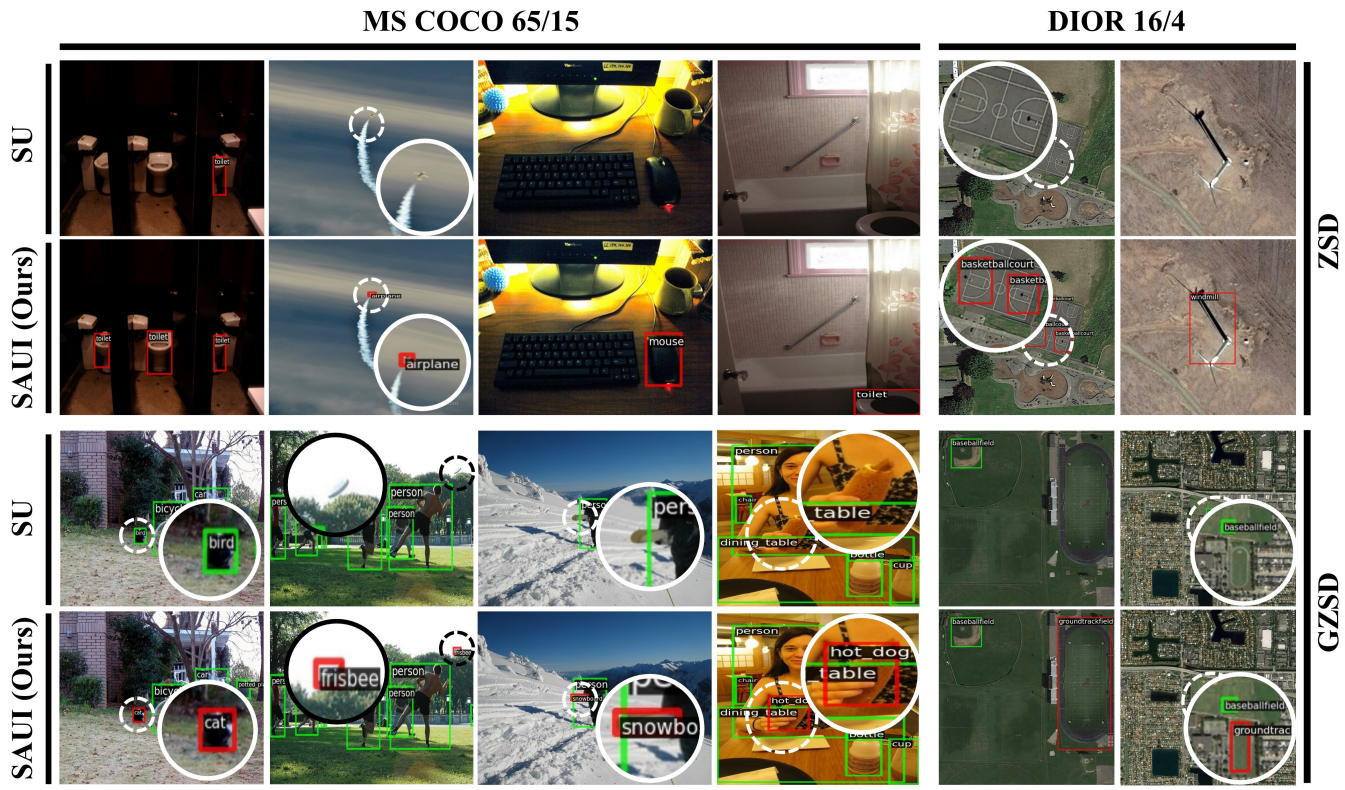


Figure 4: Comparison of detection results on COCO 65/15 and DIOR 16/4 between the reproduced SU and our SAUI. Seen-class and unseen-class objects are denoted by green and red boxes, respectively. Small objects are zoomed for better illustration.

bring in any vision knowledge of CLIP image encoder to ensure the strict compliance with zero-shot settings. For MS COCO/DIOR/PASCAL VOC, we train Faster-RCNN for 14/14/18 epochs respectively. We set α_1 in Eq. (1) to $10^{-1}/10^{-1}/10^{-2}$ and sampling radius r to $10^{-4}/10^{-4}/10^{-6}$, respectively. Besides, $\alpha_2, \alpha_3, \alpha_4$ in Eq. (1) are set to $10^{-3}, 10^{-3}, 10^{-4}$, while the temperature coefficient τ is set to 10^{-1}. The number of negative samples N is 10. For fair comparison, other settings are consistent with previous works (Hayat et al. 2020; Huang et al. 2022).

Comparison with State-of-the-arts

Performance on MS COCO. In Table 1, we make comprehensive comparisons between SAUI and current state-of-the-arts. For 48/17 split, SAUI scores the best 13.7% mAP, 66.0% recall in ZSD setting and 20.7% mAP-HM in GZSD setting. For 65/15 split, SAUI achieves remarkable results with 22.1% mAP, 77.8% recall in ZSD and 27.0% mAP-HM, 70.3% recall-HM in GZSD. Our method surpasses the state-of-the-art on 9 types of evaluation metrics (10 types in total). Particularly, SAUI obtains significant gains over the second-best RSC (Sarma et al. 2022) in ZSD mAP (from 20.1% to 22.1%) , in ZSD recall with IoU=0.5 (from 65.1% to 77.8%) and in GZSD mAP-HM (from 26.2% to 27.0%). Table 2 details the class-wise AP performance of a series of generation-based methods, in which SAUI outperforms other methods on 9 unseen classes (15 unseen classes in total). We also reproduce two previous works to obtain their multi-scale AP scores. As shown in Table 2, SAUI stands out from the crowd on detection of small objects with 9.9% AP_s, which equals almost 3 times the second-best score. Meanwhile, SAUI achieves significant increase in AP_m (from 11.0% to 14.9%) and in AP_l (from 25.2% to 28.3%). Superior multi-scale AP scores demonstrate SAUI's strong capacity in detecting small and scale-varying objects.

Performance on PASCAL VOC. As shown in Table 3, SAUI achieves 65.5% ZSD mAP and 48.0% GZSD mAP-HM. In terms of the class-wise performance, SAUI scores 93.5% mAP on *dog* and 53.7% mAP on *train*. Thus, SAUI achieves comparable results to the state-of-the-art. This is reasonable because objects in PASCAL VOC are generally prominent, central and large-scale. When dealing with this kind of objects, SAUI will degrade to vanilla methods and fail to exploit its advantages.

Performance on DIOR. Table 4 reveals the dominance of SAUI on DIOR. More specifically, SAUI achieves an outstanding performance with 16.1% mAP in ZSD setting and 11.7% unseen mAP, 17.0% mAP-HM in GZSD setting. In comparison, SAUI obtains 42% growth percentage on ZSD mAP score and surprising 178.7% growth percentage on GZSD mAP-HM score over the second-best RRFS (Huang et al. 2022), leading by a large margin. This is because objects in remote sensing images tend to be scale-varying, which brings tremendous challenge to existing ZSD meth-

Method	ZSD					GZSD		
	car	dog	sofa	train	mAP	S	U	HM
SAN	56.2	85.3	**62.6**	26.4	57.6	48.0	37.0	41.8
HRE	55.0	82.0	55.0	26.0	54.5	**62.4**	25.5	36.2
PL	**63.7**	87.2	53.2	44.1	62.1	-	-	-
BLC	43.7	86.0	60.8	30.1	55.2	58.2	22.9	32.9
SU*	59.6	92.7	62.3	45.2	64.9	-	-	-
RRFS*	60.1	93.0	59.7	49.1	**65.5**	47.1	**49.1**	**48.1**
SAUI	45.7	**93.5**	62.1	**53.7**	**65.5**	47.2	48.9	48.0

Table 3: Comparison of mAP at IoU=0.5 on PASCAL VOC dataset, including class-wise AP in ZSD setting in addition.

Method	ZSD	GZSD		
		S	U	HM
PL	0.4	4.3	0.0	0.0
BLC	1.1	6.1	0.4	0.8
SU*	10.5	30.9	2.9	5.3
RRFS*	11.3	30.9	3.4	6.1
SAUI	**16.6**	**31.3**	**11.7**	**17.1**

Table 4: Comparison of mAP at IoU=0.5 on DIOR dataset in ZSD and GZSD settings.

ods. Our SAUI can generate scale-wise diverse features for unseen classes and thus achieves a leap forward in zero-shot remote sensing object detection.

Ablation Study

Effects of different modules in SAUI. To evaluate the effect of each module, we conduct ablation experiments on DIOR dataset. We construct 5 ablation models which exclude Series-GAN (SerGAN), clip-based word embeddings (CWE), Intra-scale Outline Maintain loss ($\mathcal{L}_{\text{main}}$), Intra-scale Detail Diverage loss ($\mathcal{L}_{\text{div}}$), and Inter-scale Mutual Construct loss ($\mathcal{L}_{\text{con}}$), respectively. As shown in Table 5, Series-GAN offers a significant contribution with 4.4% and 3.2% absolute increase in ZSD mAP and GZSD mAP-HM over vanilla GAN. Model without clip word embeddings drops 1.1% mAP and 0.5% mAP-HM. In addition, $\mathcal{L}_{\text{main}}/\mathcal{L}_{\text{div}}/\mathcal{L}_{\text{con}}$ each accounts for $0.7\%/0.9\%/1.1\%$ and $0.4\%/0.7\%/0.8\%$ absolute increase in mAP and mAP-HM.

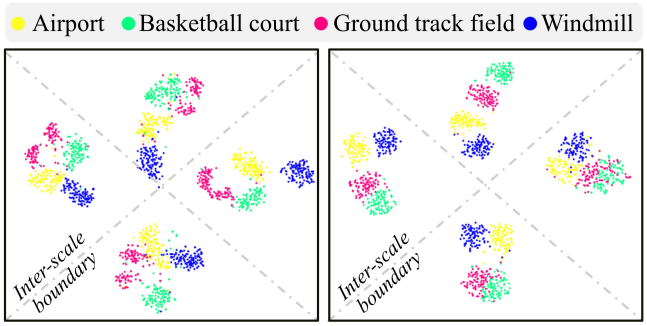
Qualitative Results

Detection Results. Fig. 4 illustrates some detection results of the reproduced SU (Hayat et al. 2020) and our SAUI on MS COCO and DIOR. We can observe from these examples that SAUI checks for the missing objects and corrects the mistakes compared to the baseline model. This is a typical phenomenon throughout our experiments and is apparent for small objects (*e.g. frisbee*) in particular, which further substantiates the multi-scale AP results in Table 2. Moreover, SAUI also shows robustness when objects of the same class vary intensively across images (*e.g. ground track field*).

Imagination Visualization. To visualize the performance of the Series-GAN synthesizer with multi-scale contrastive imagination, we apply t-SNE to real multi-scale features and synthesized multi-scale features of unseen-class objects in Fig. 5. The real unseen features show the diversity of different scale-views as we can see a clear boundary between every two scale-views. As shown, our SAUI imitates the real distributions well, with clear inter-scale boundarys in the whole visual space as well as intra-scale separability and diversity in every local scale-view space.

Method	ZSD	GZSD		
		S	U	HM
SAUI w/o SerGAN	12.2	29.5	9.1	13.9
SAUI w/o CWE	15.5	29.7	11.5	16.6
SAUI w/o $\mathcal{L}_{\text{main}}$	15.9	29.5	11.7	16.7
SAUI w/o $\mathcal{L}_{\text{div}}$	15.7	29.4	11.4	16.4
SAUI w/o $\mathcal{L}_{\text{con}}$	15.5	29.4	11.3	16.3
SAUI (full)	**16.6**	**31.3**	**11.7**	**17.1**

Table 5: Ablation studies for different modules of SAUI on DIOR dataset. "SerGAN" denotes Series-GAN structure, "CWE" denotes clip word embedding.



(a) Real unseen-class features **(b) Synthesized unseen-class features by SAUI**

Figure 5: The t-SNE visualization of (a) real and (b) synthesized unseen-class features by SAUI. Our method successfully imitates the inter-scale construction along with the intra-scale separability and diversity.

Conclusion

In this paper, we propose a novel Scale-Aware Unseen Imagineer (SAUI) that first leverages scale information to facilitate zero-shot object detection. Specifically, we utilize a coarse-to-fine extractor to capture visual features through multiple scale-views and construct a Series-GAN synthesizer to imagine scale-diverse visual features of unseen-class objects. Moreover, we apply sequence-to-sequence training strategy to the Series-GAN synthesizer and propose intra-scale and inter-scale components to stabilize, maintain and diverge the multi-scale synthesis. Extensive and comprehensive experiments conducted on three different datasets demonstrate the effectiveness of the scale-awareness motivation and the superiority of SAUI to the current state-of-the-arts, especially for scale-varying and small objects.

Acknowledgments

This work was supported by National Key R&D Program of China No. 2022ZD0117103, National Natural Science Foundation of China under Grant 62302384, 62172326 and 62137002, the MOE Innovation Research Team No. IRT17R86, China University Innovation Fund No. 2021FNA04003, China Postdoctoral Science Foundation under Grant 2023M742790, Fundamental Research Funds for the Central Universities under Grant xpt012023022, and the Project of China Knowledge Centre for Engineering Science and Technology.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the International Conference on Machine Learning*, 214–223.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-Shot Object Detection. In *Proceedings of the European Conference on Computer Vision*, 397–414. Springer.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision*, 213–229. Springer.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2018. Zero-Shot Object Detection by Hybrid Region Embedding. In *Proceedings of the British Machine Vision Conference*, 56.

Denton, E. L.; Chintala, S.; Szlam, A.; and Fergus, R. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 1486–1494.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal Of Computer Vision*, 88(2): 303–338.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1440–1448.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Han, Y.; Roig, G.; Geiger, G.; and Poggio, T. 2017. Is the Human Visual System Invariant to Translation and Scale? In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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 *Proceedings of the Asian Conference on Computer Vision*, 155–170. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7612–7621.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5830–5840.

Khan, S. H.; Hayat, M.; Zamir, S. W.; Shen, J.; and Shao, L. 2019. Striking the Right Balance With Uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 103–112.

Li, H.; Mei, J.; Zhou, J.; and Hu, Y. 2023. Zero-shot Object Detection Based on Dynamic Semantic Vectors. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 9267–9273.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *Journal of Photogrammetry and Remote Sensing*, 296–307.

Li, Z.; Yao, L.; Zhang, X.; Wang, X.; Kanhere, S. S.; and Zhang, H. 2019. Zero-Shot Object Detection with Textual Descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8690–8697.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 936–944.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Liu, W.; Chen, H.; Ma, Y.; Wang, J.; and Zheng, N. 2022. Transformer-Based Zero-Shot Detection via Contrastive Learning. In *Proceedings of the Artificial Intelligence Applications and Innovations*, 316–327. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9992–10002.

Nercessian, S. C.; Panetta, K. A.; and Agaian, S. S. 2013. Non-linear direct multi-scale image enhancement based on the luminance and contrast masking characteristics of the human visual system. *IEEE Transactions on image processing*, 22(9): 3549–3561.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Rahman, S.; Khan, S. H.; Barnes, N.; et al. 2020. Improved Visual-Semantic Alignment for Zero-Shot Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11932–11939.

Rahman, S.; Khan, S. H.; and Porikli, F. 2018. Zero-Shot Object Detection: Learning to Simultaneously Recognize and Localize Novel Concepts. In *Proceedings of the Asian Conference on Computer Vision*, 547–563. Springer.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Sarma, S.; Kumar, S.; Sur, A.; et al. 2022. Resolving Semantic Confusions for Improved Zero-Shot Detection. In *Proceedings of the British Machine Vision Conference*, 347.

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*.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12967–12974.

Zheng, Y.; Huang, R.; Han, C.; Huang, X.; and Cui, L. 2020. Background Learnable Cascade for Zero-Shot Object Detection. In *Proceedings of the Asian Conference on Computer Vision*, 107–123. Springer.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations*.