OBJECT DETECTION WITH MINIMAL SUPERVISION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BERKAN DEMIREL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

JANUARY 2023

Approval of the thesis:

OBJECT DETECTION WITH MINIMAL SUPERVISION

submitted by **BERKAN DEMIREL** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering**

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Supervisor, **Computer Engineering, METU**

Assoc. Prof. Dr. Nazlı İkizler Cinbiş
Co-supervisor, **Computer Engineering, Hacettepe University**

Examining Committee Members:

Prof. Dr. Pınar Karagöz
Computer Engineering, METU

Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Computer Engineering, METU

Prof. Dr. Pınar Duygulu Şahin
Computer Engineering, Hacettepe University

Assist. Prof. Dr. Ayşegül Dündar
Computer Engineering, Bilkent University

Assist. Prof. Dr. Emre Akbaş
Computer Engineering, METU

Date:18.01.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: BERKAN DEMİREL

Signature :

ABSTRACT

OBJECT DETECTION WITH MINIMAL SUPERVISION

Demirel, Berkan
Ph.D., Department of Computer Engineering
Supervisor: Assist. Prof. Dr. Ramazan Gökberk Cinbiş
Co-Supervisor: Assoc. Prof. Dr. Nazlı İkizler Cinbiş

January 2023, 152 pages

Object detection is considered one of the most challenging problems in computer vision since it requires correctly predicting both the object classes and their locations. In the literature, object detection approaches are usually trained in a fully-supervised manner, with a large amount of annotated data for all classes. Since data annotation is costly in terms of both time and labor, there are also alternative object detection methods, such as weakly supervised or mixed supervised learning to reduce these costs in the literature. In this thesis, our focus is handling object detection problem with minimum supervision. In this context, we first define a difficult scenario namely *zero-shot object detection* (ZSD), where no visual training data is available for some of the target object classes. Secondly, we focus on the *few-shot object detection* (FSOD) problem and propose the novel meta-tuning principle. In the ZSD problem, we propose an approach that uses visual class embeddings and convex combinations of semantic embeddings in the classification part of single-stage object detectors. Following the proposed method, we focus on using more informative word embeddings, background modeling, and potential applications for ZSD methods. We first analyze the use of embedding vectors in deep models since these vectors are an essential knowledge source for *zero-shot learning* (ZSL), and we propose a novel approach that transforms semantically meaningful word vectors into visually meaningful ones. We show that using the proposed visually meaningful word embedding vectors obtain state-of-the-art results in the *zero-shot classification* (ZSC) problem. Then, we propose the first attempt to handle the background modeling in ZSD using a novel textual attention mechanism. Finally, we introduce a new problem within the scope of ZSD applications, which we call zero-shot image captioning (ZSIC), where the input images may consist of unseen object instances. The proposed ZSIC method use template-based sentence generators and fills the empty visual template slots with object proposals obtained from ZSD methods. In this context, we also propose a new evaluation metric called *V-METEOR* to evaluate the caption qualities more accurately for the ZSIC problem. In this thesis, we also focus on the FSOD problem and propose the meta-tuning principle, which allows us to model interpretable loss functions/data augmentation magnitudes in few-shot settings. Meta-tuning allows learning inductive biases that boost FSOD as an intermediate learning step using episodic learning. With the proposed RL-based meta-tuning approach, we model the loss function parameters and augmentation magnitudes, and obtain state-of-the-art results in the FSOD problem.

Keywords: Zero-shot, Few-shot, Object Detection, Image Captioning, Meta-tuning

ÖZ

ASGARİ DENETİM İLE NESNE TESPİTİ

Demirel, Berkan
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Ramazan Gökberk Cinbiş
Ortak Tez Yöneticisi: Doç. Dr. Nazlı İkizler Cinbiş

Ocak 2023 , 152 sayfa

Nesne tespiti, hem nesne sınıflarının hem de konumlarının doğru bir şekilde tespit edilmesini gerektirdiğinden, bilgisayarlı görü alanındaki en zorlu problemlerden biri olarak kabul edilir. Literatürde önerilen nesne tespit yaklaşımları, genellikle tüm sınıflar için büyük miktarda etiketli verinin olduğu tam denetimli yöntemlerle eğitilmektedir. Veri etiketleme hem zaman hem de işçilik açısından maliyetli olduğundan literatürde bu maliyetleri azaltmak için zayıf denetimli veya karma-denetimli gibi alternatif nesne tespit yöntemleri de bulunmaktadır. Bu tezde odak noktamız, nesne tespit problemini asgari denetim ile ele almaktır. Bu bağlamda, önce bazı hedef nesne sınıfları için hiçbir görsel eğitim verisinin bulunmadığı *sıfir-atım nesne tespiti* (SAT) adlı zor bir senaryo tanımlıyoruz. Ardından, *az-atım nesne tespit* (AANT) problemine odaklanıyoruz ve meta-uyarlama ilkesini öneriyoruz. SAT probleminde, tek aşamalı nesne tespit yöntemlerinin sınıflandırma bölümünde görsel sınıf katışımlarını ve semantik katışımların dışbükey kombinasyonlarını kullanan bir yaklaşım öneriyoruz. Önerdiğimiz yöntemin ardından, daha bilgilendirici kelime katışımları, arka plan modelleme ve ZSD yöntemleri için potansiyel uygulamalara odaklanıyoruz. Bu vektörler, *sıfir-atım öğrenme*

(SAÖ) için temel bir bilgi kaynağı olduğundan, önce derin modellerde katışım vektörlerinin kullanımını analiz ediyoruz ve semantik olarak anlamlı kelime vektörlerini görsel olarak anlamlı hale dönüştüren yeni bir yaklaşım öneriyoruz. Önerilen görsel olarak anlamlı kelime katışım vektörlerini kullanmanın, *sıfır-atım sınıflandırma* (SAS) probleminde en iyi sonuçlar elde ettiğini gösteriyoruz. Ardından, hazırladığımız özgün metinsel ilgi mekanizmasını kullanarak SAT problemindeki arka plan modellemesini ele almak için literatürdeki ilk yöntemi öneriyoruz. Son olarak, SAT uygulamaları kapsamında, girdi görüntülerinin görünmeyen nesne örneklerinden oluşabileceği *sıfır-atım görüntü altyazılama* (SAGA) adını verdiğimiz yeni bir problem sunuyoruz. Önerilen SAGA yöntemi, şablon tabanlı cümle oluşturucuları kullanır ve boş görsel şablon alanlarını SAT yöntemlerinden elde edilen nesne önerileriyle doldurur. Bu kapsamda, SAGA problemi için üretilen altyazı kalitesini daha doğru bir şekilde değerlendirebilmek amacıyla *V-METEOR* adlı yeni bir değerlendirme metriği de öneriyoruz. Bu tezde, ayrıca AANT problemine odaklanıyoruz ve az-atım ayarlarında yorumlanabilir kayıp fonksiyonlarını/veri artırma büyüklükleri modellememizi sağlayan meta-uyarlama ilkesini öneriyoruz. Meta-uyarlama, epizodik öğrenmeyi kullanarak bir ara öğrenme adımı olarak AANT sonuçlarını iyileştirecek tümevarımsal önyargıların öğrenilmesine olanak sağlar. Önerilen RL tabanlı meta-uyarlama yaklaşımıyla, kayıp fonksiyon parametrelerini ve büyütme büyüklüklerini modelliyoruz ve AANT probleminde en iyi sonuçları elde ediyoruz.

Anahtar Kelimeler: Sıfır-atım, Az-atım, Nesne Tespiti, Görüntü Altyazılama, Meta-uyarlama

To my family

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisors Dr. Ramazan Gökberk Cinbiş and Dr. Nazlı İkizler-Cinbiş for their continuous guidance and support throughout my Ph.D. study. Without their guidance, expertise, and wisdom, it would have been impossible to write this thesis. I feel very lucky to be working with them not only during this doctoral thesis but also in my MS thesis and other academic studies.

In addition to my thesis advisors, I would also like to thank my thesis committee members, Dr. Pınar Karagöz, Dr. Pınar Duygulu Şahin, Dr. Ayşegül Dündar, and Dr. Emre Akbaş, for their time, valuable comments, and insights.

Besides, I would like to thank all my friends and colleagues for their support, help, and good wishes in my academic and personal life.

I would also like to thank TUBITAK for their support during this thesis. This thesis was supported in part by the TUBITAK Grants 116E445 and 119E597. Some of the numerical calculations reported in this thesis were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Last but not the least, I would like to thank my parents, my dear wife Hande, and other family members for their endless support, patience, and unwavering love during this process. I would like to dedicate this thesis to them.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xvi

LIST OF FIGURES xviii

LIST OF ABBREVIATIONS xxii

NOMENCLATURE xxiv

CHAPTERS

1 INTRODUCTION 1

  1.1 Motivation and Problem Definition 3

  1.2 Scope of the Thesis 4

    1.2.1 Zero-shot Object Detection 5

    1.2.2 Zero-shot Image Captioning 6

    1.2.3 Few-shot Object Detection 7

  1.3 Contributions and Novelties 8

  1.4 Publications 9

1.5 The Outline of the Thesis 10

2 RELATED WORK 13

2.1 Zero-shot Learning 13

2.1.1 Zero-shot Classification 13

2.1.2 Zero-shot Object Detection 16

2.1.3 Alternative Paradigms for Reducing the Dependency on Fully-supervised Learning 19

2.1.4 Zero-shot Image Captioning 19

2.1.5 Fine-grained Image Captioning 21

2.1.6 Evaluation Metrics for Image Captioning 22

2.2 Few-shot Learning 22

2.2.1 Few-shot Classification 22

2.2.2 Few-shot Object Detection 23

2.2.3 Automated Loss Function Discovery 25

2.2.4 AutoML for Data Augmentation 26

3 ZERO-SHOT OBJECT DETECTION BY HYBRID REGION EMBEDDING 29

3.1 Overview 29

3.2 Method 31

3.2.1 Region Scoring by Convex Combination of Class Embeddings 31

3.2.2 Region Scoring by Label Embedding 32

3.2.3 Zero-Shot Object Detection 33

3.3 Experiments 34

3.3.1 Datasets 34

3.3.2 Class Embeddings 36

3.3.3 Zero-Shot Detection on Fashion-ZSD Dataset 36

3.3.4 Zero-Shot Detection on Pascal-ZSD Dataset 37

3.4 Chapter Summary 40

3.5 Fashion-ZSD Dataset 40

4 ZERO-SHOT OBJECT DETECTION IN THE WILD 43

4.1 Overview 43

4.2 Method 47

4.2.1 Semantic Attention 48

4.2.2 Zero-shot Detector 49

4.2.3 Inference for Unseen Classes 50

4.3 Experiments 50

4.3.1 Implementation Details 50

4.3.2 Main Results 51

4.4 Chapter Summary 53

5 LEARNING VISUALLY CONSISTENT LABEL EMBEDDINGS FOR ZERO-SHOT LEARNING 55

5.1 Overview 55

5.2 Method 57

5.2.1 Visually Meaningful Vector Space Word Vectors 57

5.2.2 Label Embedding 58

5.3 Experiments 59

5.3.1 Datasets 59

5.3.2 Implementation Details 59

5.3.3 Experimental Results 60

5.4 Chapter Summary 63

6 CAPTION GENERATION ON SCENES WITH SEEN AND UNSEEN OBJECT CATEGORIES 65

6.1 Overview 65

6.2 Method 68

6.2.1 Main zero-shot detection model 68

6.2.2 Generalized zero-shot detection extensions 69

6.2.3 Zero-shot captioning model 71

6.2.4 Measuring zero-shot captioning quality 72

6.3 Experiments 74

6.3.1 Experimental setup 75

6.3.2 Generalized zero-shot object detection 75

6.3.3 Zero-Shot image captioning 80

6.3.4 V-METEOR experiments 85

6.3.5 Additional analyses 87

6.3.5.1 Diagnosing errors 87

6.3.5.2 Impact of using similarity embeddings 88

6.4 Chapter Summary 90

7 META-TUNING LOSS FUNCTIONS AND DATA AUGMENTATION FOR FEW-SHOT OBJECT DETECTION 91

7.1 Introduction 91

7.2 Method 94

7.2.1 Meta-tuning loss functions 96

7.2.2 Meta-tuning augmentations 97

7.2.3 Meta-tuning procedure 97

7.3 Experiments 98

7.3.1 Main results 100

7.3.2 Ablation studies 104

7.4 Chapter Summary 108

7.5 More on Meta-tuning 108

7.5.1 Proxy task class splits 108

7.5.2 Algorithm 109

7.5.3 Additional Experimental Results 110

7.5.4 Implementation and runtime 112

8 CONCLUSIONS 121

REFERENCES 125

LIST OF TABLES

TABLES

Table 3.1 ZSD performances of proposed hybrid method on Fashion-ZSD dataset. 37

Table 3.2 ZSD performances of proposed label embedding (LE), convex combination (CC) and hybrid (H) methods on Pascal-ZSD dataset. 39

Table 4.1 Experimental results on MS-COCO (65/15), Pascal VOC (16/4) and Pascal VOC (17/3) datasets. 51

Table 5.1 Zero-shot classification performance of proposed method on AwA and aPaY datasets. 61

Table 5.2 Comparison of the related ZSL literature. 62

Table 6.1 mAP results on MS-COCO dataset with GZSD (65/15) settings. . . 77

Table 6.2 mAP results on MS-COCO dataset in the 65/15 GZSD setting. . . . 80

Table 6.3 Our results on ZSD and GZSD (72/8). 81

Table 6.4 Zero-shot captioning results with comparison to captioning models involving visually fully-supervised models. 83

Table 6.5 V-METEOR comparison results. 85

Table 6.6 mAP results on MS-COCO dataset with GZSD (65/15) settings. . . 90

Table 7.1 FSOD and G-FSOD results on Pascal VOC and MS-COCO datasets for MPSR baseline method. 99

Table 7.2 FSOD (mAP) and G-FSOD (HM of the base and novel class mAPs) results on Pascal VOC. 102

Table 7.3 Comparison of Meta-ScaledDynamic results to the fine-tuning based GFSOD methods on the MS-COCO dataset. 104

Table 7.4 Evaluation of meta-tuning details. 105

Table 7.5 Low-shot (1-shot, 2-shot and 3-shot) experiments on MS-COCO dataset with novel classes. 107

Table 7.6 Proxy task class splits for Pascal VOC. 109

Table 7.7 Comparison to fine-tuning based FSOD methods on the Pascal VOC dataset, with only novel classes. 113

Table 7.8 Comparison to fine-tuning based G-FSOD methods on the Pascal VOC dataset, with both base and novel classes. 114

Table 7.9 Comparison to meta-learning based FSOD methods on the Pascal VOC dataset, with only novel classes. 115

Table 7.10 Comparison to meta-learning G-FSOD methods on the Pascal VOC dataset, with both base and novel classes. 116

Table 7.11 FSOD and G-FSOD results on the MS COCO dataset with novel classes. 120

LIST OF FIGURES

FIGURES

Figure 1.1 An example of the object detection problem. 1

Figure 1.2 An illustration of single-stage object detection frameworks. This framework is representative and shows that no RPN or ROI pooling layers are used in the middle stages. This figure does not fit all single-stage models. 2

Figure 1.3 An illustration of two-stage object detection frameworks. This framework is representative and does not fit all two-stage models. 3

Figure 1.4 Diagram for research within the scope of this thesis. 4

Figure 2.1 An example of label embedding approaches. The figure is taken from [1]. 14

Figure 2.2 An approach that uses purely textual descriptions for ZSL. The figure is taken from [2]. 15

Figure 2.3 An example of feature-space generative models. The figure is taken from [3]. 16

Figure 2.4 A framework for balancing prediction confidences between seen and unseen classes. The figure is taken from [4]. 17

Figure 2.5 Semantic clustering-based loss term to bring similar classes closer to each other. The figure is taken from [5]. 18

Figure 2.6 A method for transforming image classifiers into object detectors. The figure is taken from [6]. 19

Figure 2.7 A framework for dense image captioning tasks. The figure is taken from [7]. 20

Figure 2.8 Image captioning approach that uses the user intent to obtain controllable image captions. The figure is taken from [8]. 21

Figure 2.9 An example of adaptation-based FSL approaches. The figure is taken from [9]. 22

Figure 2.10 An example of meta-learning based FSOD approaches. The figure is taken from [10]. 23

Figure 2.11 An example of fine-tuning based FSOD approaches. The figure is taken from [11]. 24

Figure 2.12 An example of label embedding approaches. The figure is taken from [12]. 26

Figure 2.13 An example of automatic data augmentation methods. The figure is taken from [13]. 27

Figure 3.1 The framework of our ZSD model. 31

Figure 3.2 Sample images for generated toy Fashion-ZSD dataset. 35

Figure 3.3 Successful detection results of unseen objects on Pascal-ZSD dataset using proposed hybrid region embedding. 38

Figure 3.4 Unsuccessful detection results of unseen objects on Pascal-ZSD dataset using hybrid region embedding. 38

Figure 3.5 Randomly selected images from Fashion-ZSD dataset. 41

Figure 4.1 Comparison of most relevant ZSD models in the literature. . . . 44

Figure 4.2 An overview of the proposed network architecture. 47

Figure 4.3 Some visual ZSD results on MS-COCO dataset. (Best viewed in color.) 52

Figure 4.4 mAP values for different maximum number of object proposals for RPN. 53

Figure 5.1 We propose a zero-shot learning approach based on visually meaningful word vectors and label embedding. 56

Figure 5.2 Top-3 highest scoring images using PBT method in the AwA dataset. 60

Figure 6.1 (a) *Partial zero-shot image captioning* problem, (b) *True zero-shot image captioning* problem. 66

Figure 6.2 Our zero-shot captioning framework. 67

Figure 6.3 Summary of the proposed GZSD method. 69

Figure 6.4 α scaling factor learning process. 70

Figure 6.5 GZSD results on scenes containing various **seen** and **unseen** class instances. 78

Figure 6.6 The accuracy values of the proposed method in the GZSD test splits of MS-COCO according to different alpha scaling factors. 79

Figure 6.7 Image captioning results on images with *seen* and **unseen** class instances. 84

Figure 6.8 V-METEOR results of each unseen classes. 86

Figure 6.9 Image captioning results of NBT-baseline and our methods. . . . 87

Figure 6.10 False positive analyses for superclasses on MS-COCO. 89

Figure 7.1 The overall architecture of the meta-tuning approach. 91

Figure 7.2 Details of our meta-tuning approach. 95

Figure 7.3 Qualitative Pascal VOC results using MPSR without (first row) and with (second row) meta-tuning. 103

Figure 7.4 The dynamic temperature functions and score scaling coefficients learned by the meta-tuning process. 106

Figure 7.5 Randomly sampled *MPSR+Meta-ScaledDynamic* object detection results for the Pascal VOC dataset Split-3/10-shot experiment. . . . 117

Figure 7.6 Randomly sampled *MPSR+Meta-ScaledDynamic+Aug* object detection results for the Pascal VOC dataset Split-3/10-shot experiment. 118

Figure 7.7 Randomly sampled *DeFRCN+Meta-ScaledDynamic+Aug* object detection results for the Pascal VOC dataset Split-2/10-shot experiment. 119

LIST OF ABBREVIATIONS

AP	Average Precision
aPaY	aPascal-aYahoo dataset
AwA	Animals with Attributes dataset
CNN	Convolutional Neural Network
FSOD	Few-shot Object Detection
GZSL	Generalized Zero-shot Learning
GZSD	Generalized Zero-shot Object Detection
GFSOD	Generalized Few-shot Object Detection
RL	Reinforcement Learning
MLP	Multilayer Perceptron
HRE	Hybrid Region Embedding
Pascal VOC	Pascal Visual Object Classes Challenge
GAP	Global Average Pooling
BCE	Binary Cross-entropy
RPN	Region Proposal Network
ROI	Region of Interest
mAP	Mean Average Precision
MS COCO	Microsoft Common Objects in Context
HM	Harmonic Mean
SA	Semantic Attention
PL	Polarity Loss
ZSL	Zero-shot Learning
ZSC	Zero-shot Classification
ZSIC	Zero-shot Image Captioning

ZSD	Zero-shot Object Detection
CC	Convex Combination
LE	Label Embedding
L_1	Least Absolute Deviation Loss
IoU	Intersection Over Union
NBT	Neural Baby Talk
IBT	Image-based Training
PBT	Predicate-based Training

NOMENCLATURE

$x \in \mathcal{X}$	Input image
$\mathcal{Y}_s$	Set of seen classes
$\mathcal{Y}_u$	Set of unseen classes
y	Annotation of a given image or bounding box
$b \in \mathcal{B}$	Candidate bounding box
(x, y, h, w)	Bounding box regression coordinates of related object region
t	Bounding box confidence score for seen classes
ϕ	Embedding vector
$f_{\text{CC}}(x, b, y)$	Function for measuring the relevance of the label y for a given candidate bounding box b
$\eta(y)$	d_e dimensional embedding vector, such as word embeddings of class names or class-wise attribute indicator vectors
ϕ_{CC}	Image region embedding
$f_{\text{LE}}(x, b, y)$	Label embedding driven scoring function
$\phi_{\text{LE}}(x, b)$	Deep convolutional neural network that maps the image region b of image x to the space of class embeddings
$f \in R^{\frac{h}{s}x\frac{w}{s}xC}$	Visual feature map
h	Height of input images
w	Width of input images
C	Number of feature channels
P	Candidate object proposal
s	Parameter for how much the spatial size of input images reduced
R_p	Same-sized ROIs
N	Number of proposals

K	Spatial dimension of each ROI
J	Joint visual features and corresponding word embeddings
$\odot$	Hadamard product
$\Delta(\bullet, \bullet)$	Margin function, indicates pairwise discrepancy value for each given training classes
$s(\bullet, \bullet)$	Compatibility function, measures the relevance between a pair of class and a set of posterior-probability weighted attribute
λ	Regularization weight
Φ	Transformation matrix
ξ	Slack variables
Θ	Visual descriptors of given input image
W	Matrix encodes textual and visual data to assign unseen test classes to correct class label
SxS	Prediction grid of size
$f(x, c, i)$	Prediction score corresponding to the class c and cell i
c	Corresponding class
i	Corresponding cell
$\Psi(c)$	c-th class embedding
$\Omega(x, i)$	Predicted cell embedding
α	Unseen/novel class scaling coefficient
ℓ_h	Uncertainty calibration loss
τ	Softmax temperature coefficient
$p_u(\bullet)$	$f(x, c, i)$-driven unseen class likelihoods
$\varphi(c)$	c-th class name's word embedding
ω	Model parameters
q	Ground-truth caption
$p(q\|x; \omega)$	Conditional caption likelihood
r_t	Latent variable to represent the specific image region

P	Unigram precision value
R	Unigram recall value
p	Penalty term for evaluating the overall sentence compatibility
F_{mean}	Form of harmonic mean
F^{v}_{mean}	F_{mean} of visual entities
F^{n}_{mean}	F_{mean} of non-visual entities
m	Number of unigrams in both reference and generated captions
w_t	Number of unigrams in the candidate caption
w_r	Number of unigrams in the reference caption
s_c	Number of maximally long matching subsequences
u_m	Number of mapped unigrams
C_b	Base classes
C_n	Novel classes
k	Number of training images for novel classes
$\ell_{cls}(x, y)$	MPSR classification loss term
N_{ROI}	Number of ROIs in an image
$f(x, y)$	Corresponding class y prediction score in image x
ρ_τ	Temperature scalar
f_ρ	Dynamic temperature function
$\rho = (\rho_a, \rho_b, \rho_c)$	3-tuple of polynomial coefficients
n	Normalized fine-tuning iteration index
ρ_{aug}	Shared magnitude parameter
T	Proxy task
$C_{\text{p-base}}$	Proxy base classes
$C_{\text{p-novel}}$	Proxy novel classes
$D_{\text{p-pretrain}}$	Dataset for pre-training the meta-tuning model containing $C_{\text{p-base}}$-only samples

$D_{\text{p-support}}$	Dataset for finetuning the meta-tuning model containing $C_{\text{p-base}} \cup C_{\text{p-novel}}$ samples
$D_{\text{p-query}}$	Dataset for evaluating the generalized FSOD performance of fine-tuning model during meta-tuning containing $C_{\text{p-base}} \cup C_{\text{p-novel}}$ samples
$\mathcal{N}(\mu_j, \sigma^2)$	Gaussian distribution for modeling each ρ_j parameter
μ	Mean value for gaussian distributions
σ	Standard deviation value for gaussian distributions
$p(\rho; \mu, \sigma)$	Gaussian probability density function
η	RL learning rate
$R(\rho)$	Normalized reward function

CHAPTER 1

INTRODUCTION

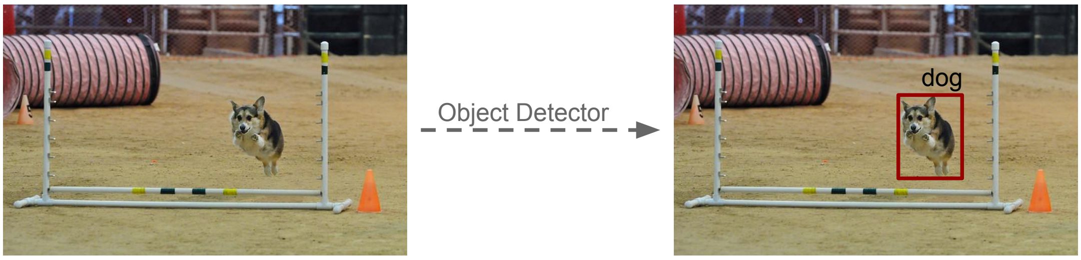


Figure 1.1: An example of the object detection problem.

Object detection, which aims to predict the classes and locations of the object instances in the images, has gained tremendous momentum with the high performance of deep learning models [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Object detection problem offers a wide range of usage scenarios in the real world, from home robotics to self-driving cars. Due to the potential impact and the difficulty of the problem, object detection has been studied for a long time as a trending field in computer vision. Object detection, which gained great momentum with deep learning methods, is previously handled with methods that use handcrafted designs such as Viola-Jones [26], HOG Detector [27], or Deformable Part Models [28] approaches. Deep learning-based approaches have been at the forefront in recent years, as in other computer vision research fields.

Most deep learning-based object detection methods can be classified as follows:

- **Single-stage approaches:** Single-stage approaches [14, 15, 21] include methods that perform classification and regression tasks simultaneously at the end of architecture using dense sampling without generating candidate object regions

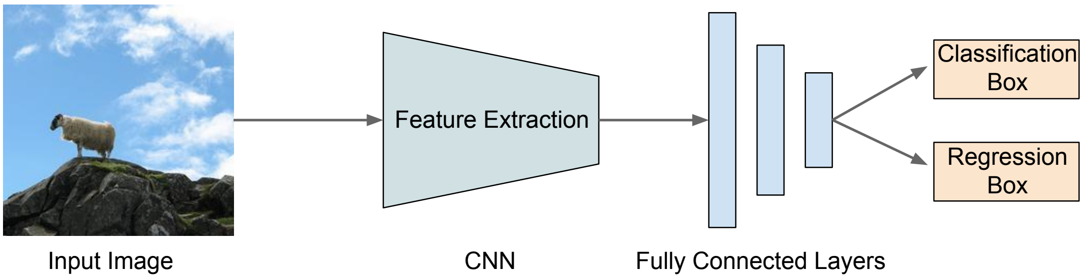


Figure 1.2: An illustration of single-stage object detection frameworks. This framework is representative and shows that no RPN or ROI pooling layers are used in the middle stages. This figure does not fit all single-stage models.

in intermediate steps. These methods are architecturally simpler than two-stage approaches as they do not include layers such as RPN (region proposal network), ROI (region of interest) Pooling. Thanks to their simple architecture, they are generally faster than equivalent two-stage models, but until recently, they achieve typically lower accuracy than two-stage approaches. However, thanks to recent methods, competitive single-stage models are being proposed in terms of both efficiency and accuracy [29]. Figure 1.2 shows an illustration of a single-stage object detection process.

- **Two-stage approaches:** Two-stage object detection approaches [16, 17, 23] consist of the generation of candidate object regions with RPN, then classification and box refinement steps on these candidate regions after ROI pooling. Since two-stage detectors include two different stages, they are slower and computationally more expensive than single-stage methods. We show an illustration of two-stage object detection in Figure 1.3.

To sum up, both single-stage and two-stage object detection approaches have their own advantages and disadvantages. Which method will be used will also vary according to the desired accuracy and speed requirements.

Recently, transformer models have gained dominance in the field of natural language processing [30, 31, 32, 33, 34], and have also achieved successful results in the computer vision [35, 36, 25, 37, 38]. In the object detection problem, visual backbone features are extracted using CNN-based pre-trained models in single-stage and two-

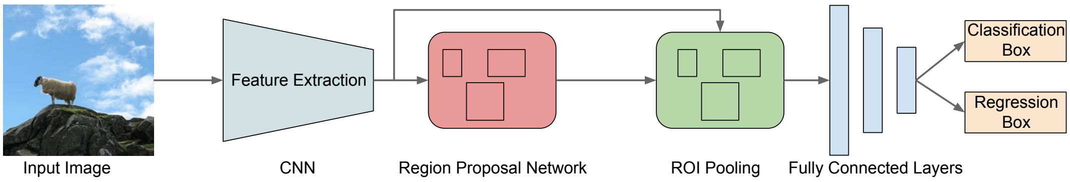


Figure 1.3: An illustration of two-stage object detection frameworks. This framework is representative and does not fit all two-stage models.

stage models until recently. Alternatively, after the adaptation of transformers to computer vision, successful results are also obtained by using transformers in backbone models [37].

1.1 Motivation and Problem Definition

Object detection approaches are usually trained in a fully-supervised manner, where there is a large amount of annotated data for all classes. In the 2014 release of the MS COCO dataset [39], one of the prominent datasets on the subject, the number of images for 80 classes in the training, validation, and test sets are 82783, 40504, and 40775, respectively. These images contain an average of 3.5 categories, and there are an average of 7.7 instances per image. Despite the fact that the state-of-the-art in object detection is impressive [37, 40, 29], these statistics show that object detectors still have problems with semantic scalability. As the object detection methods use fully-supervised training schemes, there is a need to collect a large amount of data, and this is very costly in terms of time and labor due to its nature. Due to this bottleneck, there has been growing interest in techniques that can lower the cost of data labeling, such as weakly supervised [41, 42] or mixed supervised learning [43].

In this thesis, we focus on approaches that can minimize data collection costs as an alternative to supervised methods. In this context, we first define the zero-shot object detection (ZSD) problem. The ZSD problem (see Section 1.2.1 for more details) can be defined as the adaptation of the ZSL paradigm to the object detection models. The rationale for describing the ZSD problem is that the ZSL motivation (see Section 1.2.2 for more details) is more suited to the object detection problem:

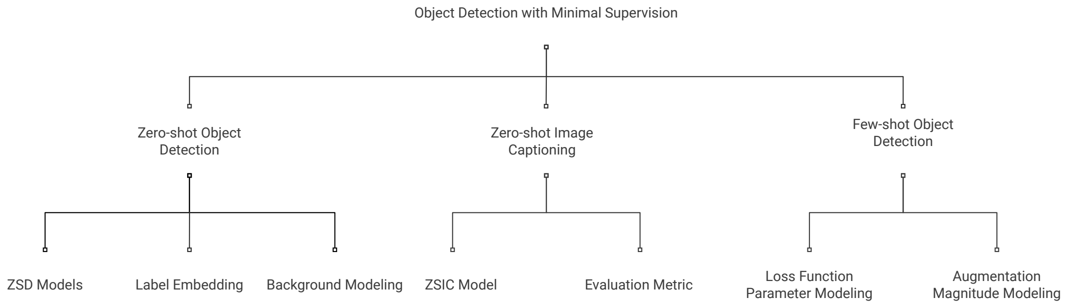

Object Detection with Minimal Supervision
Zero-shot Object Detection
Zero-shot Image Captioning
Few-shot Object Detection
ZSD Models
Label Embedding
Background Modeling
ZSIC Model
Evaluation Metric
Loss Function Parameter Modeling
Augmentation Magnitude Modeling

Figure 1.4: Diagram for research within the scope of this thesis.

object detection requires more detailed and dense data annotation than the image classification task. In this thesis, inspired by our ZSD approach, we also define the true zero-shot image captioning (ZSIC) problem for which the ZSL paradigm is also appropriate. The purpose of this problem is to adapt the ZSL paradigm to the image captioning problem and to minimize the data collection costs for the image captioning problem. Finally, within the scope of this thesis, we also focus on the few-shot object detection (FSOD) problem (see Section 1.2.3 for more details), which is an alternative minimal supervision approach. In this problem, unlike the ZSL paradigm, there are a small but non-zero number of instances for each novel class category.

1.2 Scope of the Thesis

Within the scope of this thesis, we focus on object detection with minimal supervision to handle semantic scalability problem from different perspectives. In this context, we define the ZSD problem, which is a more challenging scenario than the existing alternatives. We also focus on various topics (*e.g.* model proposal, label embedding, background modeling, evaluation, and applications) related to ZSD. In this thesis, we also propose the true ZSIC problem as a continuation of the ZSD. Finally, we define the RL-based meta-tuning concept for FSOD, which is another alternative approach, and model the loss function parameters and augmentation magnitudes statically or dynamically. Research within the scope of the thesis can be examined through Figure 1.4.

1.2.1 Zero-shot Object Detection

Zero-shot learning (ZSL) aims to enable the recognition of unseen classes which have no visual data during the training stage. Typically, this is accomplished by transferring knowledge (*e.g.* word embeddings [44], class hierarchies [45], and attributes [46]) from seen to unseen classes. The approaches used in the ZSL problem generally include using textual class similarities [44] and mapping textual and visual information in a common embedding space [1, 47, 48]. The main objective of these methods is to lower the data annotation burden for the classification task and to make an attempt to handle the scalability problem.

In this thesis, the extension of the ZSL to the object detection problem is one of the key objectives. The main reasons for this are respectively: i) the cost of data collection and hence the scalability problem is more intense for object detection, ii) ZSD is a more realistic scenario than ZSL in the context of the real-world application. Class information of objects is labeled in the image classification problem; however, in the object detection problem, the object locations are also annotated in addition to the classes, making data annotation activities more intense and costly. There are also potential real-world uses for ZSD in robotics and autonomous vehicle technologies. In this thesis, we propose the ZSD problem to detect unseen classes due to the reasons and motivation mentioned above. Our research within the scope of ZSD covers the following topics and related analyzes:

1. **Model for the ZSD Problem.** We propose our first ZSD approach based on the fusion of two widely used zero-shot image classification methods on a single-stage object detection framework (*i.e.* YOLO [15]): i) label embedding-based classification [49] and (ii) convex combination of class embeddings [44]. To be more precise, we suggest a hybrid model with two components. The first component uses the detection scores of the object detector to embed object candidates into the class embedding space. Moreover, the second component discovers a direct mapping from image regions to the class embedding space. The cosine similarities between both combinations of these region embeddings and correct class embeddings are calculated to obtain detection scores of candidate regions.

2. **Label Embedding.** In our ZSD approach, we directly employ class embeddings to detect object classes. Hence, we need to conduct ZSL research on label embedding in order to determine the relationship between word embedding vectors and visual features. Since prior knowledge is the main resource of knowledge transfer, the performance of a ZSL method is highly dependent on the quality of word embeddings. In the ZSL domain, label embedding methods typically use word vectors [50, 51] of class names that are obtained from textual data [52, 53, 54], and primarily reflect semantic relations, so we believe that word vectors should also reflect the visual aspects since ZSL is a computer vision problem.
3. **Background Modeling.** ZSD training methods employ the same training schemes as recent fully supervised object detection methods, collecting low ground-truth overlap regions as negative samples. In this scenario, if the training images contain instances of unlabeled classes, it is difficult to generate candidate boxes for these unlabeled classes during the inference stage since these class instances might be learned as background regions. To avoid this problem, current ZSD methods, to the best of our knowledge, eliminate images containing instances of unseen classes from training sets [55, 56]. However, this situation is against the nature of the ZSD problem in two aspects and creates a dilemma: (i) images consisting of selected unseen classes should be known and discarded during the training time, (ii) the image level annotations of the selected unseen classes are already known during training. Thus, if instances of unlabeled unseen classes are available in the training set, it is possible for these classes to be modeled as background regions by object detection models.

1.2.2 Zero-shot Image Captioning

ZSL has emerged as a promising alternative for overcoming practical limitations in collecting labeled image datasets and building image classifiers with extremely large object vocabulary. Similarly, *zero-shot image captioning* (ZSIC) aims to develop strategies for circumventing the data annotation bottleneck in the image captioning problem. However, we find no earlier work that is specifically designed to handle image captioning problem in a truly zero-shot setting.

1. **Model for the ZSIC Problem.** Recent works on ZSIC [57, 58] focus solely on the language domain, assuming the availability of a pre-trained fully-supervised object detector that covers all object classes of interest. We refer to these methods as *partial zero-shot image captioning*. In light of these observations, we propose the problem of true zero-shot captioning, in which test images contain instances of unseen object categories with no supervised visual or textual examples, in addition to the seen categories. We believe that this change constitutes a more direct problem definition towards (i) developing semantically scalable captioning methods, and, (ii) evaluating captioning approaches in a realistic setting where not all object classes have training examples.

2. **New Evaluation Metric.** In this thesis, we observe that using existing metrics for the evaluation of ZSIC models causes some deficiencies. Existing evaluation metrics are designed to measure the quality of the sentence by giving the same amount of penalty to all words without distinguishing between visual and non-visual words. In this case, even if there are unsuccessful ZSD models for the ZSIC problem, high evaluation metric scores can be obtained if other non-visual words are in the groundtruth sentence structure. In order to handle this situation, we propose V-METEOR as a new metric that distinguishes between visual and non-visual words based on the existing METEOR [59] metric within the scope of this thesis.

1.2.3 Few-shot Object Detection

The aim of the *few-shot object detection* (FSOD) is to build object detection models for *novel* classes that have few labeled training images by transferring knowledge from the *base* classes that have a large number of labeled images. The purpose of the closely related Generalized-FSOD (G-FSOD) problem is to build few-shot detection models that work well on both base and novel classes. Meta-learning and fine-tuning procedures are two types of FSOD methods. Although meta-learning-based methods [60, 61, 62, 63, 64, 10, 65, 66, 67, 68] are widely employed in FSOD research, numerous fine-tuning based approaches have lately reported competitive results [11, 69, 70, 71, 72, 73, 74, 75].

The core concept of meta-learning approaches is to design and train dedicated meta-models that map given few train samples to novel class detection, or to learn easily adaptable models [76] in a MAML [77] fashion. In contrast, however, fine-tuning based methods tackle the problem as a typical transfer learning problem and apply the general purpose supervised training techniques, *i.e.* regularized loss minimization via gradient-based optimization, to adapt a pre-trained model to few-shot classes. In the scope of this thesis, we focus on using meta-learning concepts to tune the loss functions and augmentations used in the fine-tuning based FSOD models, which we call *meta-tuning*:

1. **Loss Function Augmentation.** According to an analysis on the FSOD problem, there is a problem in correctly classifying the obtained candidate regions rather than finding possible candidate regions for novel classes [69]. Hence, we think it would be more appropriate to focus on the loss terms related to classification and decide to model the temperature parameter. In this context, we propose an RL-based model to learn the optimal temperature parameter of the loss functions both statically and dynamically.
2. **Data Augmentation Magnitudes.** Data augmentation is an important factor affecting success in object detection. Using the appropriate augmentation list in optimal magnitudes can contribute positively to success. In this context, we observe that photometric augmentations are important for FSOD. We then use our proposed meta-tuning approach to model their magnitudes.

1.3 Contributions and Novelties

Our contributions are as follows:

- We define a novel zero-shot setting for detecting objects of unseen classes, and propose a novel hybrid method to handle this newly defined task. This hybrid method uses a convex combination of class embeddings, and label embedding based classification together.
- We introduce two new benchmarks for evaluating ZSD approaches based on Fashion-MNIST [78] and Pascal VOC [79] datasets. The first of these benchmark

datasets, Fashion-ZSD, is a newly created dataset using Fashion-MNIST images. The other dataset created within the scope of Pascal VOC is obtained as new splits by determining the new seen and unseen classes.

- We obtain visually meaningful class name embeddings by learning to associate corresponding attribute combinations and class names, and use them within a label embedding framework.
- We propose another novel ZSD approach that incorporates a probability scaling scheme for the generalized zero-shot object detection (GZSD) problem.
- We examine the background modeling problem for ZSD and propose a first attempt to handle it, to the best of our knowledge. In this context, we propose a semantic attention mechanism and use customizable feature maps according to the input side information.
- We analyze the GZSD failure patterns, which are all directly relevant to object detection and image captioning qualities. In this context, we used four failure patterns: localization errors, confusion with background, class confusion within superclass members, and class confusion across superclasses.
- We define a paradigm for generating captions of unseen classes, which is called true zero-shot image captioning. We evaluate several caption evaluation metrics and discuss their suitability for the zero-shot image captioning scenario.
- We propose the V-METEOR metric and use this new metric for more detailed analyses of the ZSIC models. This metric explicitly measures the joint visual or non-visual accuracy of a sentence.
- We define the meta-tuning paradigm to tune the loss functions and augmentations to be used in the fine-tuning stage for FSOD. We reduce the computational costs compared to other methods that aim to discover loss terms from scratch by defining meta-tuning over well-designed loss terms and an augmentation list.

1.4 Publications

Problem definitions and contributions within the scope of the thesis mentioned above are discussed extensively in our publications/submissions below. Moreover, the materials used in this writing are obtained from our following published/submitted papers in the scope of the thesis:

1. **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Zero-Shot Object Detection by Hybrid Region Embedding", British Machine Vision Conference (BMVC), August 2018.
2. **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Learning Visually Consistent Label Embeddings for Zero-Shot Learning", IEEE International Conference on Image Processing (ICIP), September 2019. **(Oral Presentation)**.
3. **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Image Captioning with Unseen Objects", British Machine Vision Conference (BMVC), September 2019. **(Spotlight Presentation).**
4. **Berkan Demirel**, Ramazan Gokberk Cinbis, "Caption Generation on Scenes with Seen and Unseen Object Categories", Image and Vision Computing (IMAVIS), 2022.
5. **Berkan Demirel**, Orkun Öztürk, Mehmet Can Baytekin, Ramazan Gokberk Cinbis, "Zero-shot Object Detection in the Wild". **(Submitted)**.
6. **Berkan Demirel**, Orhun Buğra Baran, Ramazan Gokberk Cinbis, "Meta-tuning Loss Functions and Data Augmentation for Few-shot Object Detection". **(Submitted)**.

1.5 The Outline of the Thesis

The organization of this thesis is as follows. We describe related works related to the thesis topic in detail in Chapter 2. We explain our initial work on ZSD with details in Chapter 3. Then, we share the details of the background modeling problem that we analyzed within the scope of the ZSD problem in Chapter 4. We present our study on word embeddings that we use in our ZSD models in Chapter 5. Then, we describe the ZSIC problem, which we introduced by derivation from the ZSD problem, in Chapter 6. This section also contains the details of the V-METEOR metric that we have presented to the literature. Moreover, we describe our proposed meta-tuning mechanism for the FSOD problem with detail in Chapter 7. Finally, we summarize our findings on zero/few-shot object detection problems and other related topics in Chapter 8.

CHAPTER 2

RELATED WORK

In this section, we will share the related works mentioned in section 1. These related works are obtained by elaborating the relevant studies mentioned in our papers (see Section 1.4) and adding recent papers in the literature.

2.1 Zero-shot Learning

This section provides an overview of recent developments on zero-shot image classification, zero-shot object detection, and zero-shot image captioning.

2.1.1 Zero-shot Classification

Early work on ZSC focused on directly using attribute-based probabilistic models for transferring knowledge from seen to unseen classes [80]. Further works explore other knowledge transfer mediums and predictive models, *e.g.* [1, 53, 46, 81, 82, 83, 84, 85, 86, 2, 48, 87, 88, 89, 90]. Akata *et al.* [1] propose a discriminatively learned compatibility model over image features and attribute-based class embeddings (see Figure 2.1). Akata *et al.* [53] suggest the use of class hierarchies and distributed word representations of class names as alternatives to handcrafted attributes. Frome *et al.* [84] use convolutional neural network architectures for mapping visual features into a rich semantic embedding space. Song *et al.* [85] propose a transductive learning (QFSL) method to learn unbiased embedding space since embedding spaces often have strong bias problem. Besides, visual-semantic discrepancy problem is observed when using textual side information. In this context, Demirel *et al.* [46] use attribute

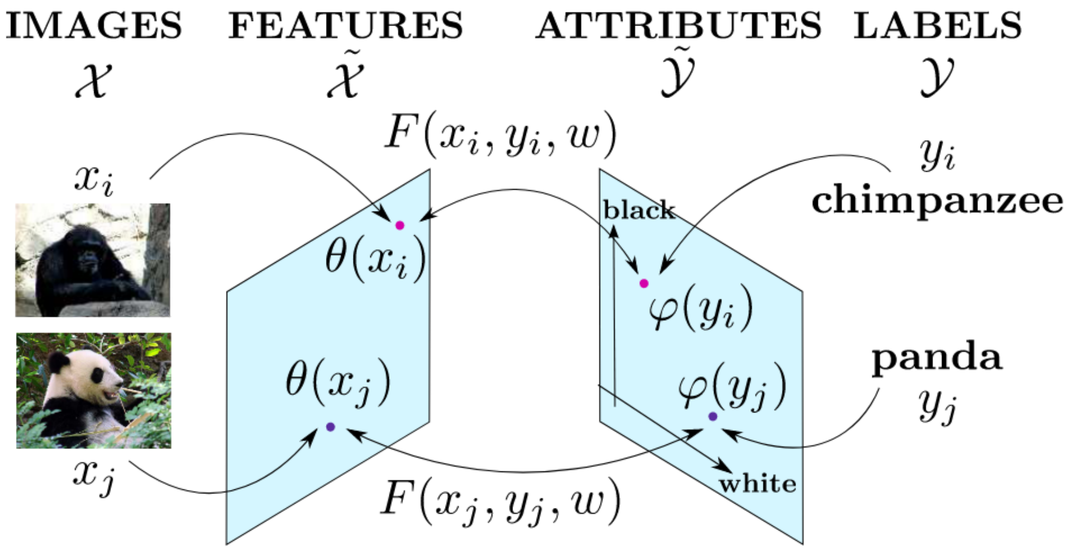


Figure 2.1: An example of label embedding approaches. The figure is taken from [1].

information as an intermediate layer to learn more generalizable distributed word representations.

Norouzi *et al.* [86] use convex combination of the semantic embedding vectors directly without learning any semantic space. Elhoseiny *et al.* [2] handle zero-shot learning problem with purely textual descriptions. They define a constrained optimization formula that combine regression and knowledge transfer functions with additional constraints (see Figure 2.2). Ba *et al.* [48] use MLP in their text pipeline to learn classifier weights of CNN in the image pipeline to handle zero-shot fine-grained object classification. The defined MLP network generates a list of pseudo-attributes for each visual category by utilizing raw texts acquired from Wikipedia articles. Other notable approaches include synthesized classifiers [81], semantic autoencoders [82], hierarchy graphs [83], diffusion regularization [87], attribute regression [88] and latent space encoding [89]. Feng *et al.* [90] propose an adversarial training mechanism for domain adaptation and disentangling visual features. A comparative survey of discriminative ZSL models can be found in Xian *et al.* [91], which introduces the problem of *generalized zero-shot learning* (GZSL) problem in an image classification context. [92] can also be followed for surveys for side information in the ZSL topic.

Alternatively, the development of generative models that can synthesize training examples of unseen classes has received significant interest in recent years, *e.g.* [3,

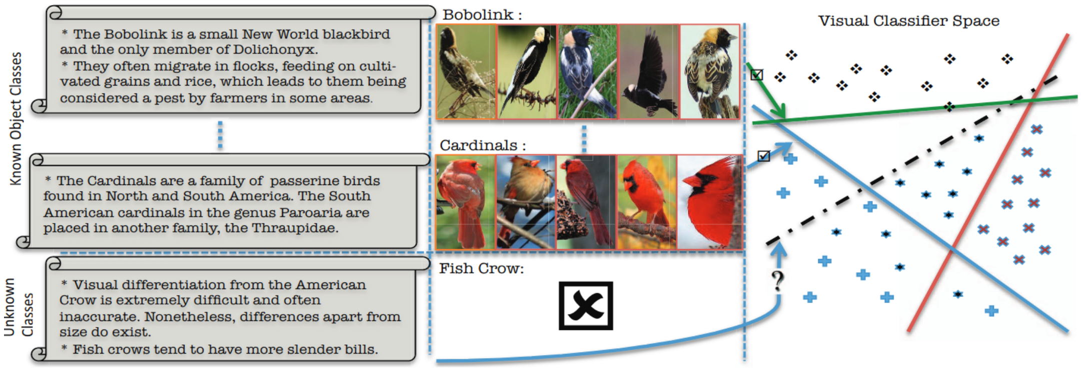

Known Object Classes
* The Bobolink is a small New World blackbird and the only member of Dolichonyx.
* They often migrate in flocks, feeding on cultivated grains and rice, which leads to them being considered a pest by farmers in some areas.
Bobolink :
* The Cardinals are a family of passerine birds found in North and South America. The South American cardinals in the genus Paroaria are placed in another family, the Thraupidae.
Cardinals :
Unknown Classes
* Visual differentiation from the American Crow is extremely difficult and often inaccurate. Nonetheless, differences apart from size do exist.
* Fish crows tend to have more slender bills.
Fish Crow:
Visual Classifier Space
?

Figure 2.2: An approach that uses purely textual descriptions for ZSL. The figure is taken from [2].

93, 94, 95, 96, 97, 98, 99, 100, 101]. These methods aim to build class embedding conditional models that can be used to generate synthetic training samples for unseen classes, therefore, they can be considered learned augmentation techniques. Data generating GZSL approaches are attractive for naturally addressing the seen class bias problem, at the cost of typically being computationally more demanding and formulationally more complex than purely discriminative approaches. In this context, Bucher *et al.* [3] propose to build feature-space generative models as one of the first works in this direction (see Figure 2.3). Felix *et al.* [93] propose to use data reconstruction for model regularization, based on multi-modal cycle consistency loss term. Mishra *et al.* [94] propose a conditional Variational Autoencoder (VAE) [102, 103] based model, and Xian *et al.* [95] improves conditional VAEs via adversarial training. Zhu *et al.* [96] propose to learn textual description conditional generative models. Li *et al.* [97] utilize conditional Wasserstein GANs. Sariyildiz and Cinbis [98] propose gradient matching loss to improve the quality of the generated samples. Chen *et al.* [99] propose a unified feature refinement network to improve the visual-semantic mapping of classes. Elhoseiny and Elfeki [104] work on the incorporation of losses that directly aim to increase sample variations. Chen *et al.* [100] suggest a disentangling model to distinguish semantically consistent and unrelated feature vectors. Su *et al.* [101] use two different autoencoders to obtain separate modalities for visual and semantic features in a common latent space. Li *et al.* [105] decompose seen attributes to their main attribute components and synthesizes new attributes.

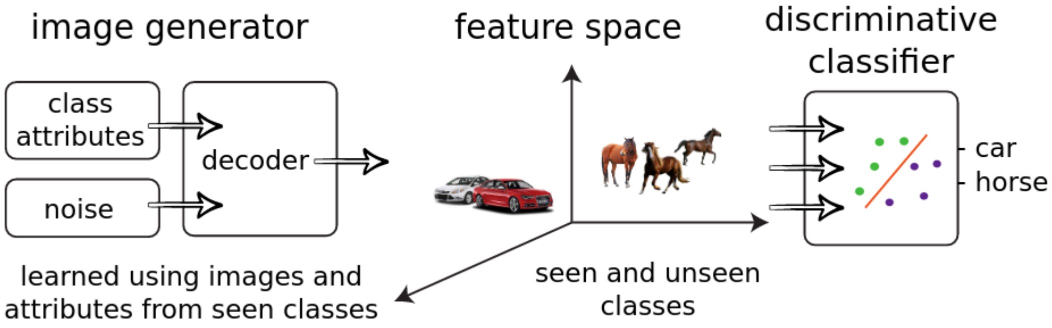


Figure 2.3: An example of feature-space generative models. The figure is taken from [3].

One of the challenges in GZSL is keeping the seen and unseen class scores comparable. In particular, discriminatively learned classification models, trained over seen class samples, tend to yield higher confidence scores for seen classes even on the test samples of unseen classes. A prominent idea in addressing this problem is reducing the prediction bias towards seen classes. For this purpose, Liu *et al.* [4] propose to increase unseen class prediction confidence by minimizing the entropy of unseen class scores during training (see Figure 2.4). Jian *et al.* [106] promote higher confidence scores for the *familiar* unseen classes during training based on unseen-to-seen class similarity estimates. Chao *et al.* [107] use an empirically chosen seen class score scaling coefficient.

Our label embedding approach for the ZSL problem is similar to [46]. Unlike [46], however, we construct our final ZSL model using the image-to-class associations measured by a label embedding classifier instead of relying directly on the attribute-to-class associations in the transformed word embedding space.

2.1.2 Zero-shot Object Detection

The most recent object detection methods can be categorized into the following two groups: (i) regression-based approaches, and, (ii) region proposal-based approaches. Regression-based approaches generate all candidate detection scores and positions jointly in a single step, using a single convolutional network [24, 15, 14, 21, 108, 109].

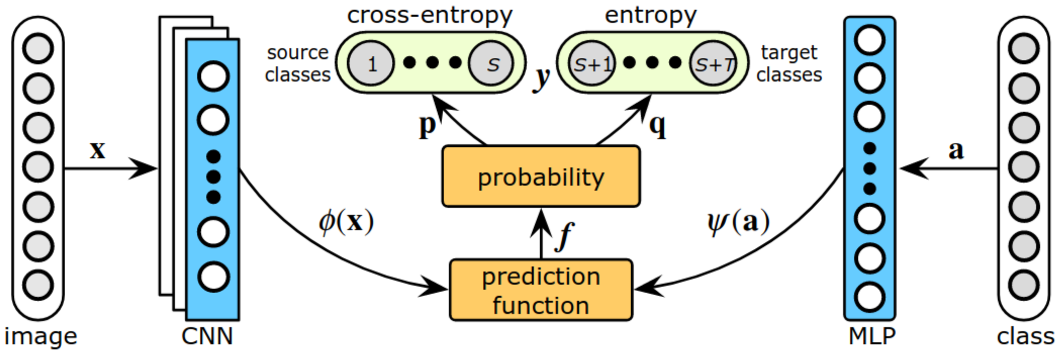


Figure 2.4: A framework for balancing prediction confidences between seen and unseen classes. The figure is taken from [4].

Region proposal-based approaches instead first generate region proposals, then classify (and update) each region proposal [110, 111, 23, 20, 17, 18, 16].

ZSD is a relatively new problem, pioneered by our first work and [5, 112, 113]. These studies were published on similar dates in the same year, and subsequently, other ZSD studies [114, 115, 116, 55, 117, 118, 119, 120, 121, 56, 122, 123] follow these papers. These approaches typically extend supervised detection models to ZSD. Among these studies, Bansal *et al.* [112] proposes a two-step approach that first locates object proposals from low-level features [124] and then classifies the resulting candidate regions using a ZSL model. Huang *et al.* [122] propose an architecture that provides both semantic divergences for intra-classes and structure-preserving for inter-classes together. Yang *et al.* [123] provide a feature-based ZSD model that generates deep features of detectors as visual features of seen and unseen objects. Other studies suggest an end-to-end framework by modifying the existing regression-based or region proposal-based detection approaches. In this context, Rahman *et al.* [5] proposes a region proposal-based approach and uses a semantic clustering-based loss term to bring similar classes closer to each other (see Figure 2.5).

ZSD methods also aim to generate results for both seen and unseen classes at the same time. However, it is also known from the zero-shot classification problem that there is a bias towards seen classes in zero-shot learning concept. Regarding this issue, Rahman *et al.* [113] proposes a polarity loss term that is based on the focal loss approach, to tackle better alignment between visual and semantic domains. Hence,

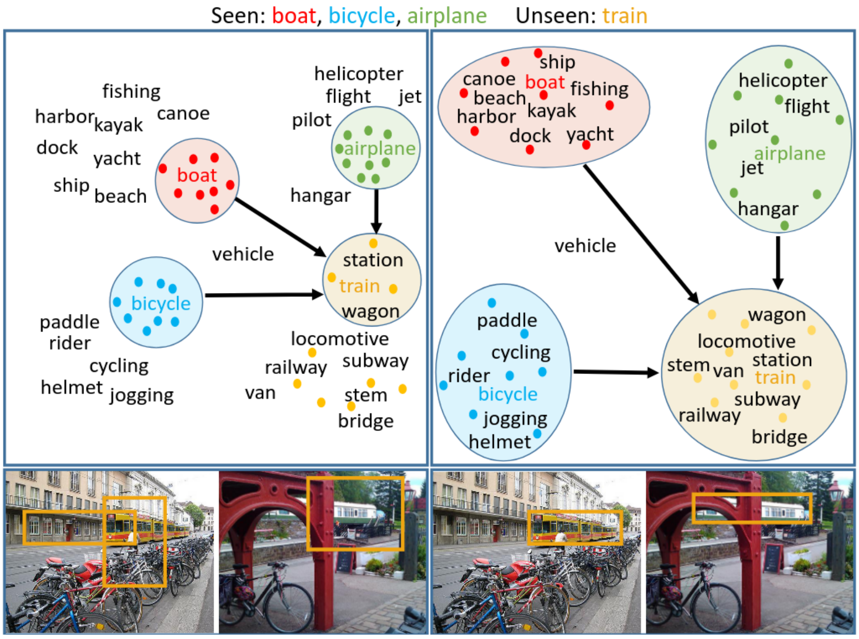


Figure 2.5: Semantic clustering-based loss term to bring similar classes closer to each other. The figure is taken from [5].

the semantic representations of visually similar classes get closer to each other. Li *et al.* [115] uses natural language descriptions of classes for ZSD. Shao *et al.* [116] focuses on the candidate proposal generation problem of unseen classes in the ZSD. Gupta *et al.* [117] learns a joint embedding space to obtain more discriminative visual and textual embeddings. Li *et al.* [118] uses a dual-path method to fuse side analogy information and knowledge transfer between the visual and textual sides. Yan *et al.* [56] uses semantics-guided network to improve conventional embeddings.

In our first approach, we propose a regression-based ZSD model that jointly incorporates convex combinations of semantic embeddings [44] and bi-linear compatibility models [1]. We also propose another ZSD component for the object detection model within the scope of the ZSIC problem. The closest detection model to the ZSD component of our ZSIC approach is our first ZSD model. Our ZSIC component differs by (i) leveraging class-to-class similarities measured in the word embedding space as class embeddings, as opposed to directly using the word embeddings, (ii) learning a class score scaling coefficient that reduces the seen class bias and improves GZSD accuracy, and (iii) exploring the use of uncertainty calibration [125] in GZSD. Finally,

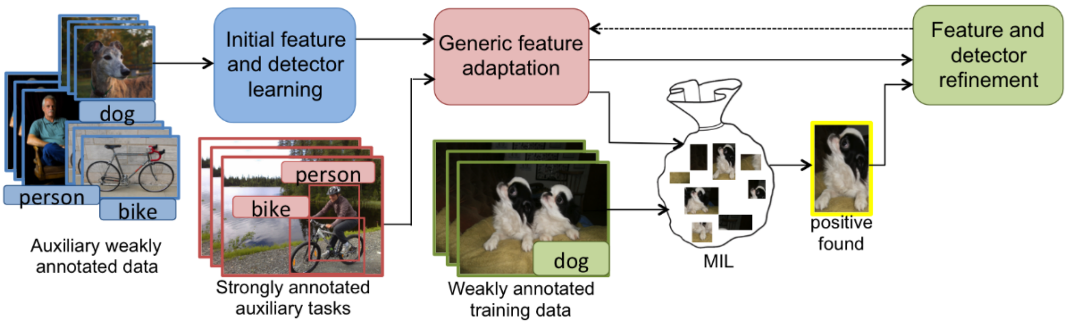

Initial feature and detector learning
Generic feature adaptation
Feature and detector refinement
dog
person
bike
Auxiliary weakly annotated data
person
bike
Strongly annotated auxiliary tasks
dog
Weakly annotated training data
MIL
positive found

Figure 2.6: A method for transforming image classifiers into object detectors. The figure is taken from [6].

we propose a third ZSD model within the scope of this thesis. This model is an approach that works on background modeling for unseen classes and it uses the novel textual attention mechanism that is proposed as a new approach by us.

2.1.3 Alternative Paradigms for Reducing the Dependency on Fully-supervised Learning

There exist alternative learning paradigms that also aim to reduce the dependency on fully-supervised training examples for object detection. To this end, methods for transforming image classifiers into object detectors (see Figure 2.6), *e.g.* [126, 6, 15], and image-level label based weakly supervised learning approaches, *e.g.* [127, 128, 129], stand out as closely related directions. However, such approaches still require labeled training images for all classes of interest, which can be a major obstacle in building models with the semantic richness needed for captioning.

2.1.4 Zero-shot Image Captioning

State-of-the-art captioning approaches are based on deep neural networks [130, 131, 132, 133, 134, 135, 57, 136, 137, 138, 139, 140]. Mainstream methods can be categorized as (i) template-based techniques [135, 141, 57] and (ii) retrieval-based ones [142, 143, 144, 133]. Template-based approaches generate templates with empty slots, and fill those slots using attributes or detected objects. Kulkarni *et al.* [135]

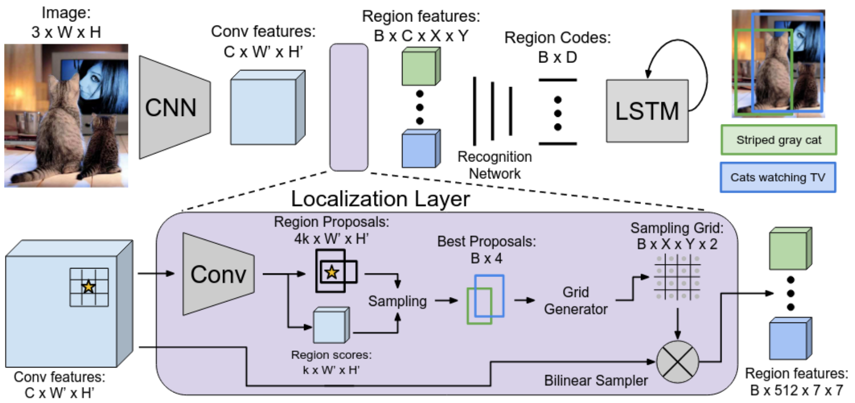


Figure 2.7: A framework for dense image captioning tasks. The figure is taken from [7].

builds conditional random field models to push tight connections between the image content and sentence generation process before filling the empty slots. Farhadi *et al.* [141] uses triplets of scene elements for filling the empty slots in generated templates. Lu *et al.* [57] uses a recurrent neural network to generate sentence templates for slot filling. Retrieval-based image captioning methods, in contrast, rely on retrieving captions from the set of training examples. More specifically, a set of training images similar to the test example are retrieved and the captioning is performed over their captions.

In this thesis, we aim to generate captions that can include classes that are not seen in the supervised training set, where retrieval-based approaches are not directly suitable. For this reason, we adopt a template-based approach that generates sentence templates and fills the visual word slots with the GZSD model predictions.

Dense captioning [7, 145, 146] appears to be similar to ZSIC, but the focus is significantly different: while dense captioning aims to generate rich descriptions, our goal in ZSIC is to achieve captioning over the novel object classes (see Figure 2.7). Some captioning methods go beyond training with fully supervised captioning data and allow learning with a captioning dataset that covers only some of the object classes plus additional supervised examples for training object detectors and/or classifiers for all classes of interest [147, 148, 58, 149, 150]. Since these methods presume that all necessary visual information can be obtained from some pre-trained object recognition

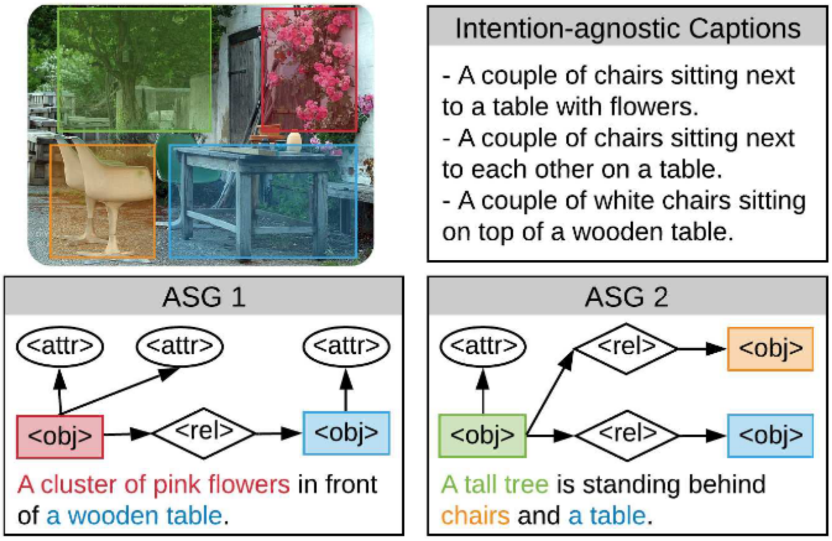

Intention-agnostic Captions
- A couple of chairs sitting next to a table with flowers.
- A couple of chairs sitting next to each other on a table.
- A couple of white chairs sitting on top of a wooden table.
ASG 1
<attr>
<attr>
<attr>
<obj>
<rel>
<obj>
A cluster of pink flowers in front of a wooden table.
ASG 2
<attr>
<rel>
<obj>
<obj>
<rel>
<obj>
A tall tree is standing behind chairs and a table.

Figure 2.8: Image captioning approach that uses the user intent to obtain controllable image captions. The figure is taken from [8].

models, we believe they cannot be seen as true ZSIC approaches.

2.1.5 Fine-grained Image Captioning

Recently, fine-grained image captioning methods are also proposed to generate richer image captions [8, 151, 152, 153]. The purpose of these methods is not to generate captions for novel objects, but to generate more descriptive captions for the classes available in the training set. Chen *et al.* [8] uses *Abstract Scene Graphs* (ASG) to obtain controllable image captions according to the user intent at the desired dense level. ASG is a graph-based scene layout to represent user intentions in the generated captions (see Figure 2.8). Khan *et al.* [151] uses *Bahdanau attention* [154] on visual features to obtain isolated image content for rich visual embeddings. Yuan *et al.* [152] proposes a gated mechanism to adjust the weights of global and local visual features. Thus, the level of detail for the caption is adjusted with respect to the visual information. Cheng *et al.* [153] adjusts attention weights of visual feature vectors and semantic feature embeddings in a decoder cell sequence to obtain rich fine-grained image captions. Unlike ZSIC, these methods do not target generating captions with objects unavailable in the training set.

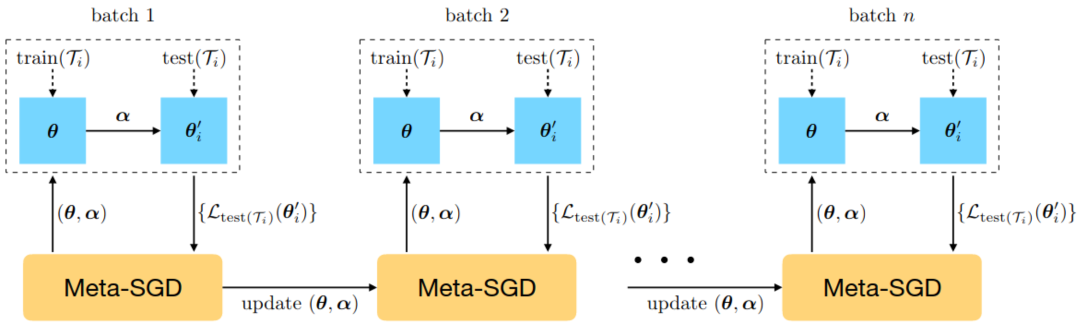

batch 1
batch 2
batch n
train($\mathcal{T}_i$)
test($\mathcal{T}_i$)
$\boldsymbol{\theta}$
$\boldsymbol{\alpha}$
$\boldsymbol{\theta}'_i$
$(\boldsymbol{\theta}, \boldsymbol{\alpha})$
$\{\mathcal{L}_{\text{test}(\mathcal{T}_i)}(\boldsymbol{\theta}'_i)\}$
Meta-SGD
update $(\boldsymbol{\theta}, \boldsymbol{\alpha})$
• • •

Figure 2.9: An example of adaptation-based FSL approaches. The figure is taken from [9].

2.1.6 Evaluation Metrics for Image Captioning

In our thesis, we additionally look into the problem of evaluating ZSIC results. The evaluation of captioning methods is not a *solved* problem. There are well-known metrics, such as METEOR [59], SPICE [155], BLEU [156], and CIDEr [157] which are widely used in image captioning. There are also recent works to improve existing captioning metrics, such as the work of Wang *et al.* [158], which incorporates uniqueness and descriptiveness aspects into SPICE. In this thesis, we aim to build a metric that allows per-class evaluation of visual and lingual caption quality so that we can explicitly evaluate the unseen and seen class captioning success, and in this context, we propose the *V-METEOR* metric.

2.2 Few-shot Learning

This section provides an overview of recent developments on few-shot image classification, few-shot object detection, automated loss function and data augmentation discovery.

2.2.1 Few-shot Classification

Most of the meta-learning approaches for few-shot learning (FSL) of classification models can be grouped as *adaptation-based* and *mapping-based* approaches. Adaptation-

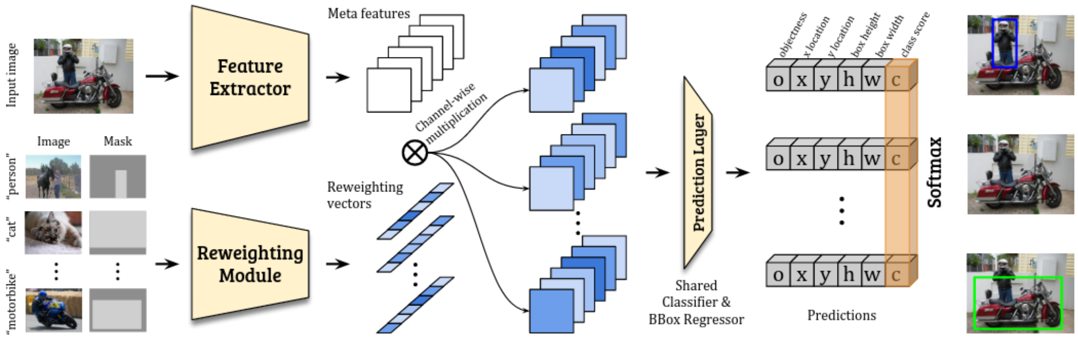


Figure 2.10: An example of meta-learning based FSOD approaches. The figure is taken from [10].

based (also called *gradient-based*) approaches aim to learn model parameters that can easily be adapted to new unseen few-shot tasks within a few model update steps (see Figure 2.9), *e.g.* [159, 9, 160, 161, 162, 163, 164]. Mapping-based approaches (also called *metric-based*) aim to bypass a gradient-descent based adaptation step, and instead learn a data-to-classifier mapping, *e.g.* [165, 166, 167, 168, 169, 170, 171, 172, 173, 174].

Some of the other notable approaches include learning to generate synthetic data for novel classes [175, 176, 177], using better feature representations [178, 179, 180, 181, 182, 183, 184] or utilizing differentiable convex solvers [185, 186]. Importantly, several works highlight that a carefully trained representation combined with simple fine-tuning or even just shallow classifiers can yield competitive or better performance than meta-learning based approaches, *e.g.* [178, 187, 188].

2.2.2 Few-shot Object Detection

The FSOD approaches can be summarized as meta-learning and fine-tuning (also called *transfer-learning*) based ones. Most meta-learning based FSOD approaches embrace formulations similar to those used in mapping-based meta-learning approaches for FSL (see Figure 2.10), *e.g.* [60, 61, 62, 63, 64, 10, 65, 66, 67, 68]. Support feature aggregation is one of the main aspects where meta-learning-based methods differ from each other. Xiao and Marlet [60] use both the differences and the channel-wise

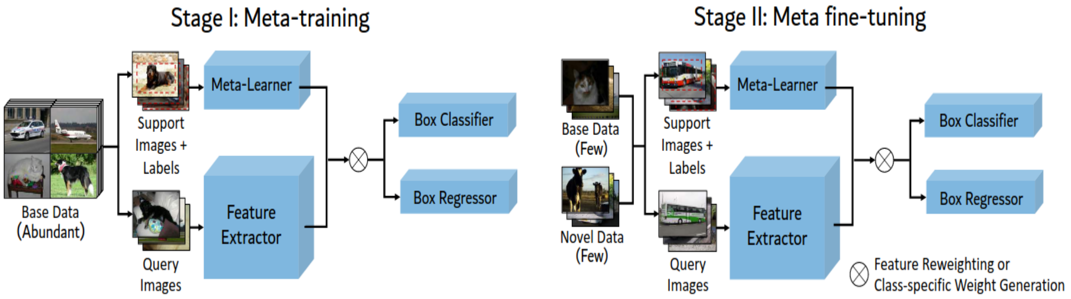


Figure 2.11: An example of fine-tuning based FSOD approaches. The figure is taken from [11].

multiplication of the features in addition to the combination of the features directly for support-query aggregation. Fan *et al.* [189] use attention blocks to make support and query features more distinguishable for base and novel object classes. Zhang *et al.* [61] use inter-class correlations to highlight important support features. Li *et al.* [62] propose to use specialized support and query features for classification and localization.

Recent efforts towards improving meta-learning based FSOD include complimentary techniques, mainly to improve loss functions, feature matching, and novel class sample usage efficiency. [62] uses class margin loss, [190] uses margin-based ranking loss, [191] uses hybrid loss which consist of focal loss, adaptive margin loss and contrastive loss. Hu *et al.* [192] perform feature matching between query and support images to use the information from the support images more effectively. Similarly, Han *et al.* [193] construct a matching network between query and support instances using heterogeneous graph convolutional networks. Li and Li [194] augment novel class samples via adding Gaussian noise. Yin *et al.* [66] decouple classification task from localization by using the proposed class-conditional architecture.

Fine-tuning-based methods typically freeze parts of a pre-trained detection network, add auxiliary detection heads, increase the novel class variances and then apply gradient descent based model update steps, unlike meta-learning-based methods that use complex episodic learning [11, 69, 70, 71, 74, 75, 195].

Wang *et al.* [11] propose a Faster-RCNN [17] based approach, where the class-agnostic region proposal network (RPN) component is kept frozen during fine-tuning (see

Figure 2.11). Sun *et al.* [69] use a similar approach and differently include FPN and RPN layers to the learnable parameter set in the same architecture. These learnable layers allow using contrastive proposal encodings that facilitate the more accurate classification of novel objects. Wu *et al.* [70] show that the scale distribution of support set tends to be imbalanced, and proposes a multi-scale positive sample refinement (MPSR) branch as an addition to the main model. Fan *et al.* [71] propose Retentive R-CNN architecture to prevent forgetting during fine-tuning for base classes. The obtained object proposals are fed into two ROI detectors responsible for base class and novel class instances. Qiao *et al.* [75] focus on decoupling network modules, and introduce a gradient decoupling layer and prototypical calibration block. Kaul *et al.* [74] extend the novel class annotations in the training set. In this context, the proposed method obtains object candidates from the base detector, and applies the box refinement step.

In the scope of this thesis, while our approach is based on fine-tuning based FSOD, we embrace meta-learn principles to optimize the loss function and augmentations to improve the fine-tuning process for FSOD, without learning a complex and over-fitting-prone meta-model. The resulting loss function and data augmentations are then utilized within the fine-tuning steps.

2.2.3 Automated Loss Function Discovery

Loss function discovery is an emerging AutoML topic towards improving the learning systems in a data-driven manner. Existing methods are mainly based on either (i) constructing the loss function directly from the basic operators [12, 196, 197] or (ii) optimizing parameterized loss functions [198, 199]. For loss construction, [12] proposes a genetic algorithm that consists of loss function verification and quality filtering modules. In this approach, the predefined proxy task eliminates divergent and poor candidate loss functions and survives the promising loss functions for other steps (see Figure 2.12). [197] uses a genetic algorithm to select candidate loss functions from a tree of simple mathematical operations, and the successful loss functions pass to other stages to mutate. [196] suggests a method to learn not only the loss function but also the whole machine learning algorithm from scratch. For loss optimization,

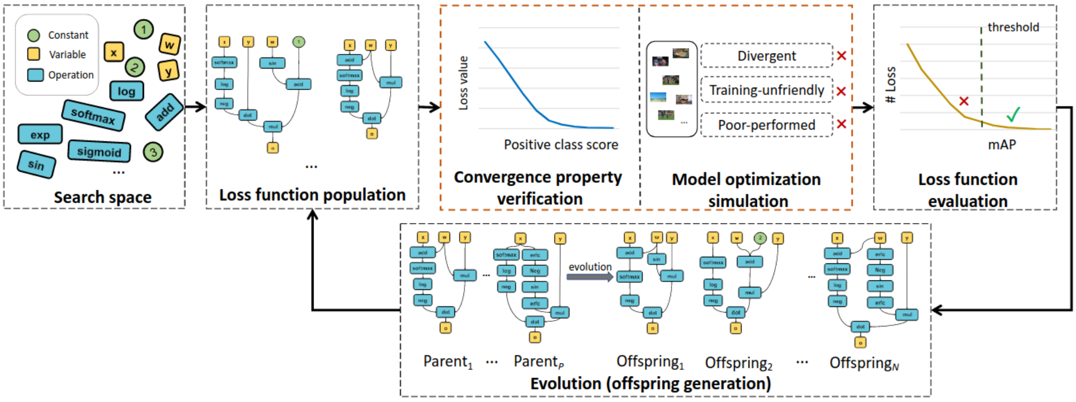


Figure 2.12: An example of label embedding approaches. The figure is taken from [12].

[198] re-analyzes the existing loss functions and presents them in a combined formula. [199] observes that the search space used in [198] can be too complex, and propose to simplify the search space via heuristics. In contrast to these works targeting supervised training scenarios, we aim to adapt loss function learning principles to the FSOD problem.

2.2.4 AutoML for Data Augmentation

A variety of automated data augmentation techniques have recently been proposed [200, 201, 13, 202]. Cubuk *et al.* [201] generate augmentation policies using reinforcement learning and a controller RNN. Ho *et al.* [200] propose a method that reduces the computational costs compared to [201] by using a population-based framework. Similarly, Lim *et al.* [13] propose a direct Bayesian method to reduce costs (see Figure 2.13). Cubuk *et al.* [202] show that the optimal augmentation magnitudes tend to be similar across transformations, and the search process can greatly be simplified by using a shared value. We follow this suggestion and use a shared magnitude across the transforms in our formulation. In contrast to these works on supervised learning, however, we focus on learning detectors with few-samples.

In summary, while loss function and augmentation discovery topics increasingly attract attention towards improving supervised training pipelines, our FSOD approach which is proposed in the scope of this thesis is the first work on learning few-sample specific

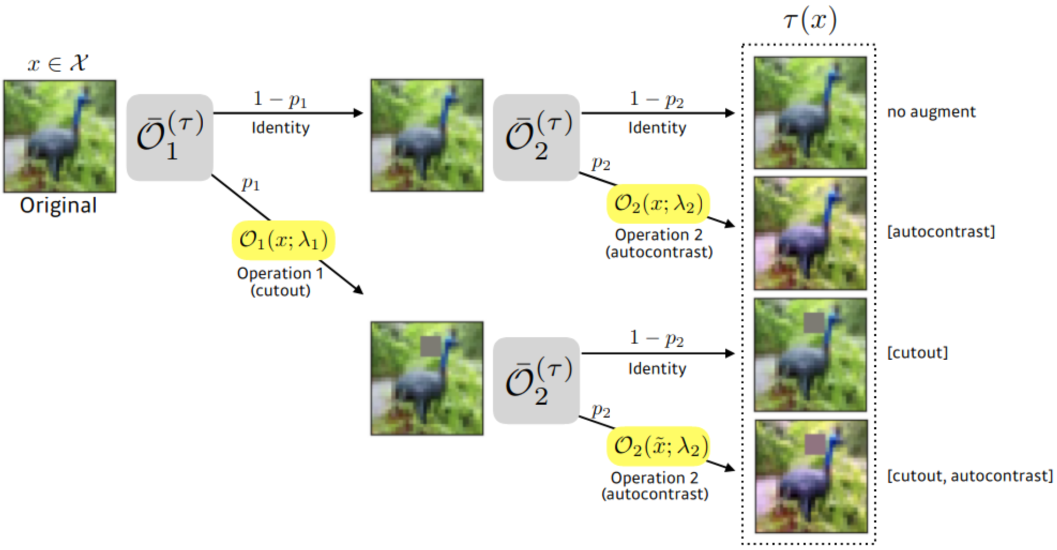


Figure 2.13: An example of automatic data augmentation methods. The figure is taken from [13].

inductive biases for fine-tuning based few-shot object detection based on meta-learning and AutoML principles, to the best of our knowledge.

CHAPTER 3

ZERO-SHOT OBJECT DETECTION BY HYBRID REGION EMBEDDING

3.1 Overview

Object detection is one of the most studied tasks in computer vision research. Previously, mainstream approaches provided only limited success despite the efforts in carefully crafting representations for object detection, *e.g.* [22]. More recently, however, CNN (convolutional neural network) based models have lead to great advances in detection speed and accuracy, *e.g.* [17, 15, 21].

While the state-of-the-art in object detection is undoubtedly impressive, object detectors still lack semantic scalability. As these approaches rely heavily on fully supervised training schemes, one needs to collect large amounts of images with bounding box annotations for each target class of interest. Due to its laborious nature, data annotation remains as a major bottleneck in semantically enriching and universalizing object detectors.

Zero-shot learning (ZSL) aims to minimize the annotation requirements by enabling recognition of unseen classes, *i.e.* those with no training examples. This is achieved by transferring knowledge from seen to unseen classes by means of auxiliary data, typically obtained easily from textual sources. Mainstream examples for such ZSL approaches include methods for mapping visual and textual information into a joint space [1, 53, 48], and, those that explicitly leverage text-driven similarities across classes [86].

The existing ZSL approaches, however, predominantly focus on classification problems. In this work, we extend this ZSL paradigm to object detection and focus on the *zero-*

shot detection (ZSD) task. Here, the goal is to recognize and localize instances of object classes with no training examples, purely based on auxiliary information that describes the class characteristics. The main motivation for studying ZSD is the observation that in most applications of ZSL, such as robotics, accurate object localization is equally important as recognition.

Our ZSD approach builds on the adaptation and combination of two mainstream approaches in zero-shot image classification: (i) convex combination of class embeddings [86], and, label embedding based classification [49]. More specifically, we propose a hybrid model that consists of two components: the first component leverages the detection scores of a supervised object detector to embed image regions into a class embedding space. The second component, on the other hand, learns a direct mapping from region pixels to the space of class embeddings. Both of these region embeddings are then converted into region detection scores by comparing their similarities with true class embeddings. Finally, we construct our zero-shot detector by integrating these two components into the the fast object detection framework YOLO [15].

We note that both components of our approach essentially provide an embedding of a given test image. Our main motivation in using them together is to employ two complementary sources of information. In particular, while the former component can be interpreted as a semantic composition method guided by class detection scores, the latter one focuses on transformation of image content into the class embedding space. Therefore, these two components are expected to better utilize semantic relations and visual cues, respectively.

In order to evaluate the effectiveness of the proposed ZSD approach, we define new benchmarks based on existing datasets. First, we create a simple ZSD dataset by composing images with multiple Fashion-MNIST [78] objects. Moreover, the Pascal VOC [203] dataset is similarly adapted to the ZSD task by defining new splits and settings. The experimental results show that our hybrid embedding approach yields promising results in both datasets.

To sum up, our main contributions in this work are as follows: (i) we define a novel zero-shot setting for detecting objects of unseen classes, (ii) we propose a novel hybrid method to handle newly defined ZSD task, (iii) we introduce two new benchmarks for

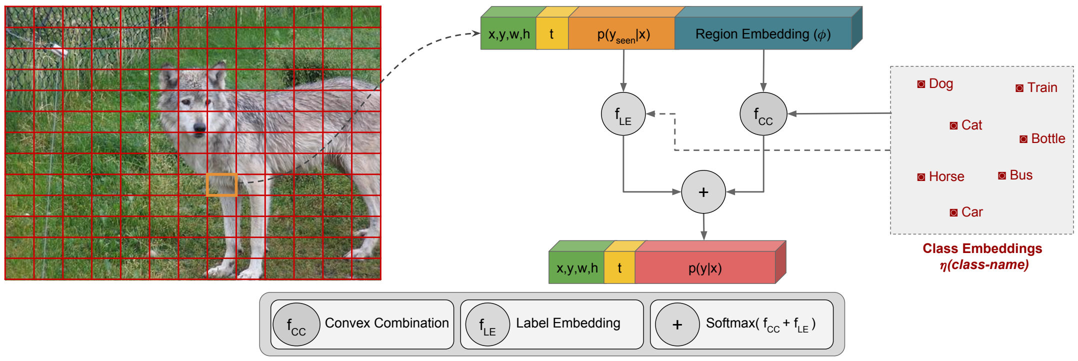


Figure 3.1: The framework of our ZSD model.

evaluating ZSD approaches based on Fashion-MNIST and VOC datasets.

3.2 Method

Our method consists of two components that (i) utilize a convex combination of class embeddings, an adaptation of the ideas from [86], and, (ii) directly learn to map regions to the space of class embeddings, by extending the label embedding approaches from zero-shot image classification [53]. Details of the model can be followed in Figure 3.1. In this model, (x, y, h, w) represents bounding box regression coordinates, t represents bounding box confidence score, $p(y_{\text{seen}}|c)$ represents initial class scores, ϕ represents embedding vector of the related region, and $p(y|x)$ represents the final zero-shot detection class probabilities.

The rest of this section explains the model details: in the first two sub-sections, we describe the convex combination and label embedding components. Then, we describe how we construct our zero-shot object detector within the YOLO detection framework.

3.2.1 Region Scoring by Convex Combination of Class Embeddings

The first component of our ZSD approach aims to semantically represent an image in the space of word vectors. More specifically, we represent a given image region (*i.e.* a bounding box) as the convex combination of training class embeddings, weighted by the class scores given by a supervised object detector of seen classes. The resulting semantic representation of the region is then utilized to estimate confidence scores for unseen classes.

This approach can be specified as follows: let $\mathcal{Y}_s$ be the set of seen classes, for which we have training images with bounding box annotations, and and let $\mathcal{Y}_u$ be the set of unseen classes, for which we have no visual training examples. Our goal is to learn a scoring function $f_{\text{CC}}(x, b, y) : \mathcal{X} \times \mathcal{B} \times \mathcal{Y} \to \mathcal{R}$ that measures the relevance of label $y \in \mathcal{Y}_s$, which can be a seen or unseen class, for a given candidate bounding box $b \in \mathcal{B}$ and the image $x \in \mathcal{X}$.

We assume that a d_e dimensional embedding vector $\eta(y)$, such as word embeddings of class names or class-wise attribute indicator vectors, is available for each class. The scoring function $f_{\text{CC}}(x, b, y)$ is then defined as the cosine similarity between the class embedding $\eta(y)$ and the image region embedding $\phi_{\text{CC}}(x, b)$:

$$f_{\text{CC}}(x, b, y) = \frac{\phi_{\text{CC}}(x, b)^{\text{T}} \eta(y)}{\|\phi_{\text{CC}}(x, b)\| \|\eta(y)\|} \tag{3.1}$$

where $\phi_{\text{CC}}(x, b)$ is defined as follows:

$$\phi_{\text{CC}}(x, b) = \frac{1}{\sum_{y \in \mathcal{Y}_s} p(y|x, b)} \sum_{y \in \mathcal{Y}_s} p(y|x, b) \eta(y) \tag{3.2}$$

Here, $p(y|x, b)$ is the class posterior probability given by the supervised object detection model. Therefore, ϕ_{CC} can simply be interpreted as a weighted sum of class embeddings, over the seen classes.

3.2.2 Region Scoring by Label Embedding

The convex combination driven scoring function fcc utilizes detection scores and embeddings of the training classes to estimate scores of zero-shot classes. In the label embedding approach, however, our goal is to directly model the compatibility between the visual features of image regions and class embeddings. For this purpose, we define the label embedding driven scoring function $f_{\text{LE}}(x, b, y) : \mathcal{X} \times \mathcal{B} \times \mathcal{Y} \to \mathcal{R}$ that measures the relevance of label $y \in \mathcal{Y}$ for a given candidate bounding box $b \in \mathcal{B}$ in an image $x \in \mathcal{X}$ as follows:

$$f_{\text{LE}}(x, b, y) = \frac{\phi_{\text{LE}}(x, b)^{\text{T}} \eta(y)}{\|\phi_{\text{LE}}(x, b)\| \|\eta(y)\|} \tag{3.3}$$

where $\phi_{\text{LE}}(x, b)$ is basically a deep convolutional neural network that maps the image region b of image x to the space of class embeddings.

We note that $f_{\text{LE}}(x, b, y)$ can equivalently be interpreted as a dot product between ℓ_2-normalized image region descriptors and class embeddings. While it is common to ℓ_2-normalize class embeddings in zero-shot image classification studies [53], we also ℓ_2-normalize the image embedding vectors. In our preliminary experiments, we have observed that this additional normalization step is beneficial for the zero-shot detection task.

We learn the $\phi_{\text{LE}}(x, b)$ network in an end-to-end fashion within our YOLO-based zero-shot detection framework, which we explain in the next section.

3.2.3 Zero-Shot Object Detection

We use the YOLO-v2 [15] architecture to construct our zero-shot object detector. The original YOLO architecture that we utilize contains a convolutional network that reduces the spatial dimensions of the input by a factor of 32 and results in a tensor of depth $k(5 + |\mathcal{Y}_s|)$, *e.g.* an input image of size $416 \times 416 \times 3$ results in a tensor of size $13 \times 13 \times k(5 + |\mathcal{Y}_s|)$. Each cell within this output tensor encodes the k detections per cell ($k = 5$ by default), and, each block of size $5 + |\mathcal{Y}_s|$ encodes one such detection. Here, for a single detection, the first 4 dimensions encode the relative bounding box coordinates, the following dimension encodes the estimated window objectness score, and the final $|\mathcal{Y}_s|$ dimensions encode class confidence scores.

To adapt YOLO architecture for the zero-shot detection task, we modify it in the following manner: we increase the final output depth from $k(5+|\mathcal{Y}_s|)$ to $k(5+|\mathcal{Y}_s|+d_e)$, where the newly added d_e dimensions per detection correspond to the $\phi_{\text{LE}}(x, b)$ output of the label embedding component of the model. In this way, the same convolutional network is shared for candidate box prediction, class prediction and class-embedding prediction purposes.

During training, the original YOLO formulation uses three separate mean-squared error based loss functions, defined over the differences between predictions and ground truth values for (i) bounding boxes, (ii) intersection-over-union values, and, (iii) classes.

For training f_{CC} defined in Eq. (3.1), the original YOLO loss function over class predictions is used as is. For training f_{LE} defined in Eq. (3.3), however, we extend the loss function by incorporating an additional loss function L_{LE}. L_{LE} basically measures correctness of the label embedding driven class predictions in a max-margin sense:

$$L_{\text{LE}}(x,b,y) = \frac{1}{|\mathcal{Y}_s|-1} \sum_{y' \in \mathcal{Y}_s \setminus \{y\}} \max\left(0, 1 - f_{\text{LE}}(x,b,y) + f_{\text{LE}}(x,b,y')\right) \tag{3.4}$$

where y is the ground-truth class corresponding to the bounding box b in input image x. Here, the goal is to ensure that at each window prediction, the label embedding based confidence score f_{LE} for the target class is larger than that of each other class. Other than this extension, we use the original YOLO training procedure, over the seen classes.

Once the network is trained, we jointly utilize the scoring functions f_{CC} and f_{LE} by computing softmax of their summations, over the classes of interest:

$$p(y|x,b) = \frac{\exp\left(f_{\text{CC}}(x,b,y) + f_{\text{LE}}(x,b,y)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(f_{\text{CC}}(x,b,y') + f_{\text{LE}}(x,b,y')\right)} \tag{3.5}$$

where $p(y|x,b)$ is the predicted posterior probability of (seen or unseen) class y given region b of image x. The final set of detections are obtained by using the non-maxima suppression procedure of YOLO over all candidate detection windows, objectness scores, and the final probabilities $p(y|x,b)$.

3.3 Experiments

In this section, we present our experimental evaluation of the proposed approach. In Section 3.3.1, we describe the ZSD datasets that we prepare and utilize. In Section 3.3.2, we explain class embeddings used in our experiments. Finally, in Section 3.3.3 and Section 3.3.4, we give the implementation details and our experimental results.

3.3.1 Datasets

We use two different datasets: Fashion-ZSD and Pascal-ZSD. We propose two new testbeds for evaluation of ZSD approaches. First, we create a synthetic dataset based on

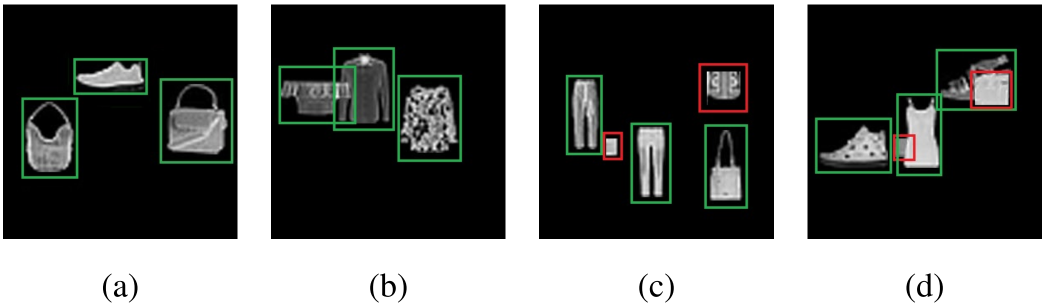

(a) (b) (c) (d)

Figure 3.2: Sample images for generated toy Fashion-ZSD dataset.

combinations of objects from the Fashion-MNIST [78] dataset. Second, we compose a new split based on existing Pascal VOC [203] benchmarks. The details of these testbeds are described below.

Fashion-ZSD. This is a toy dataset that we generate for evaluation of ZSD methods, based on the Fashion-MNIST [78] dataset. Fashion-MNIST originally consists of Zalando's article images with associated labels. This dataset contains 70,000 grayscale images of size 28x28, and 10 classes. For ZSD task, we split the dataset into two disjoint sets; seven classes are used in training and three classes are used as the unseen test classes (Table 3.1). We generate multi-object images such that there are three different objects in each image. Randomly cropped objects are utilized to create clutter regions. As shown in Figure 3.2, we consider four scenarios: from left-to-right, (a) full objects only, (b) partial occlusions, (c) clutter regions included, and (d) a scene with both partial occlusions and clutter regions. Here, ground truth object regions are shown with green and noise regions are shown in red boxes. In this dataset, 8000 images of the resulting 16333 training images are held out for validation purposes. As a result, we obtain the Fashion-ZSD dataset with 8333 training, 8000 validation and 6999 test images.

Pascal-ZSD. This is an adapted version of the Pascal VOC datasets [203]. We select 16 of the 20 classes for training and the remaining 4 classes (*i.e.* car, dog, sofa and train) for test. The *train+val* subsets of Pascal VOC 2007 and 2012 datasets are used for training classes, and the *test* subset of Pascal 2007 is used for evaluation on the unseen classes. Images containing a mixture of train and test classes are ignored.

3.3.2 Class Embeddings

For the Fashion-ZSD dataset, we generate 300-dimensional GloVe word embedding vectors [51] for each class name, using Common Crawl Data[1]. For the class names that contain multiple words, we take the average of the word vectors. For Pascal-ZSD, we use attribute annotations of aPaY dataset [204], since aPascal (aP) part of this dataset is obtained from Pascal VOC images. We average 64-dimensional indicator vectors of per-object attributes over the dataset to obtain class embeddings.

3.3.3 Zero-Shot Detection on Fashion-ZSD Dataset

In this part, we explain our ZSD experiments on Fashion-ZSD dataset. We initialize the convolutional layers of our model using the weights pre-trained on the ILSRVC12 [205] classification images. Training of our approach is completed in 10 epochs, where batch size is 32 and learning rate is 0.001. In our experiments, we first evaluate the performance of the trained network on seen training classes. According to the results presented in Table 3.1, the proposed approach obtains 91.9% mAP on the validation images with seen classes, which shows the proper training of the detection model. On the test set with unseen classes only, our proposed approach yields an mAP of 64.9%, highlighting the difficulty of zero-shot detection task even in simple settings. Here, we report class-based average precision and mean average precision (mAP) scores.

On the combinated validation and test evaluation, our method achieves 81.7% mAP. This setting is particularly interesting, as it requires recognition over both seen and unseen objects at detection time. Our result suggests that the model is able to detect objects of unseen test classes even in the presence of seen classes, without being dominated by them.

---

[1] commoncrawl.org/the-data/

Table 3.1: ZSD performances of proposed hybrid method on Fashion-ZSD dataset.

	Training Classes							Test Classes			
Test split	t-shirt	trouser	coat	sandal	shirt	sneaker	bag	pullover	dress	ankle-boot	mAP (%)
val	0.89	0.91	0.90	0.97	0.86	0.99	0.90	-	-	-	91.9
test	-	-	-	-	-	-	-	0.49	0.49	.95	64.9
val+test	0.89	0.90	0.90	0.97	0.86	0.99	0.91	0.45	0.40	0.90	81.7

3.3.4 Zero-Shot Detection on Pascal-ZSD Dataset

In this part, we explain our ZSD experiments on Pascal-ZSD dataset. Training settings of the proposed method on Pascal-ZSD dataset are same with the previous experiment, except that the number of epochs is set to 30. We present the results our approach, as well as individual performances of convex combination and label embedding components, in Table 3.2. The proposed hybrid approach yields 65.6% mAP on seen classes, 54.6% mAP on unseen classes and 52.3% mAP on the combination of seen and unseen classes. By comparing individual components of the model, we observe that convex combination (CC) outperforms label embedding (LE), and the hybrid scheme further improves the results.

The reason why the performance of the individual label embedding component is much lower can potentially be explained by the fact that the ZSD-Pascal dataset is relatively small: there are 16 classes in the training set, and this number is most probably insufficient to learn a direct mapping from visual features to class embeddings.

Qualitative results for our approach are provided in Figure 3.3. In this figure, example results of succesful detections of objects of unseen classes with various poses and sizes are shown. Additionally, example failure cases are shown on Figure 3.4. Problems in detection include missed detections, false positives, as well as misclassification of objects despite correct localization. For instance, in the second image within Figure 3.4, we see that *"picnic bench"* object is misrecognized as *"sofa"*, most probably due to relative similarity of the '*'chair"* and *"dining table"* seen classes in the embedding space.

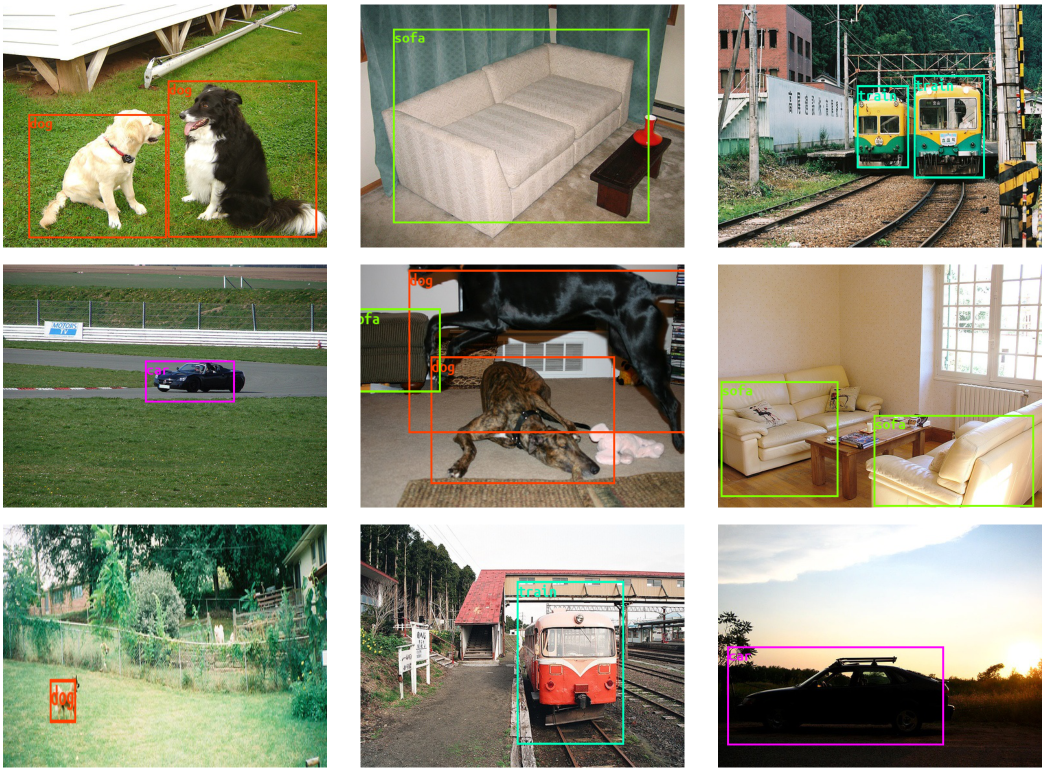


Figure 3.3: Successful detection results of unseen objects on Pascal-ZSD dataset using proposed hybrid region embedding.

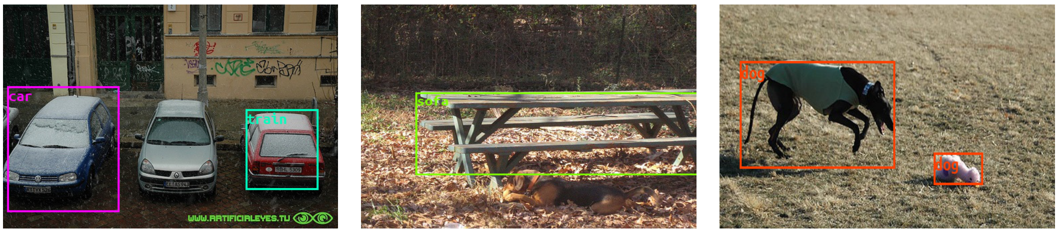


Figure 3.4: Unsuccessful detection results of unseen objects on Pascal-ZSD dataset using hybrid region embedding.

Table 3.2: ZSD performances of proposed label embedding (LE), convex combination (CC) and hybrid (H) methods on Pascal-ZSD dataset.

Method	Test split	aeroplane	bicycle	bird	boat	bottle	bus	cat	chair	cow	dining table	horse	motorbike	person	potted plant	sheep	tvmonitor	car	dog	sofa	train	mAP (%)
LE	v	0.46	0.50	0.44	0.28	0.12	0.59	0.44	0.20	0.11	0.38	0.35	0.47	0.65	0.16	0.18	0.53	-	-	-	-	36.8
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.54	0.79	0.45	0.12	47.9
	v+t	0.34	0.48	0.40	0.23	0.12	0.34	0.28	0.12	0.09	0.32	0.28	0.36	0.60	0.15	0.13	0.50	0.27	0.26	0.20	0.05	27.4
CC	v	0.69	0.74	0.72	0.63	0.43	0.83	0.73	0.43	0.43	0.66	0.78	0.80	0.75	0.41	0.62	0.75	-	-	-	-	65.0
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.60	0.85	0.44	0.27	53.8
	v+t	0.67	0.73	0.70	0.59	0.41	0.61	0.58	0.32	0.32	0.65	0.74	0.68	0.72	0.39	0.57	0.72	0.49	0.24	0.10	0.15	52.0
H	v	0.70	0.73	0.76	0.54	0.42	0.86	0.64	0.40	0.54	0.75	0.80	0.80	0.75	0.34	0.69	0.79	-	-	-	-	**65.6**
	t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.55	0.82	0.55	.26	**54.2**
	v+t	0.68	0.72	0.74	0.48	0.41	0.61	0.48	0.25	0.48	0.73	0.75	0.71	0.73	0.33	0.59	0.57	0.44	0.25	0.18	0.15	**52.3**

3.4 Chapter Summary

Localization of instances of unseen classes is as important as their recognition in various applications, such as robotics. Moreover, to overcome the annotation bottleneck, alternative methods for training object detectors are needed. To this end, in this work, we handle the problem of zero-shot detection and propose a novel hybrid method that aggregates both label embeddings and convex combinations of semantic embeddings together in a region embedding framework. By integrating these two components within an object detector backbone, detection of classes with no visual examples becomes possible. We introduce two new testbeds for evaluating ZSD approaches, and our experimental results indicate that the proposed hybrid framework is a promising step towards achieving ZSD goals.

3.5 Fashion-ZSD Dataset

In this section, we share some images (see Figure 3.5) of the toy Fashion-ZSD dataset that we generated using the Fashion-MNIST dataset [78] for the ZSD problem. As we shared in Figure 3.2, there are 4 different scenarios in our Fashion-ZSD dataset: full objects only, partial occlusions, clutter regions included, and a scene with both partial occlusions and clutter regions.

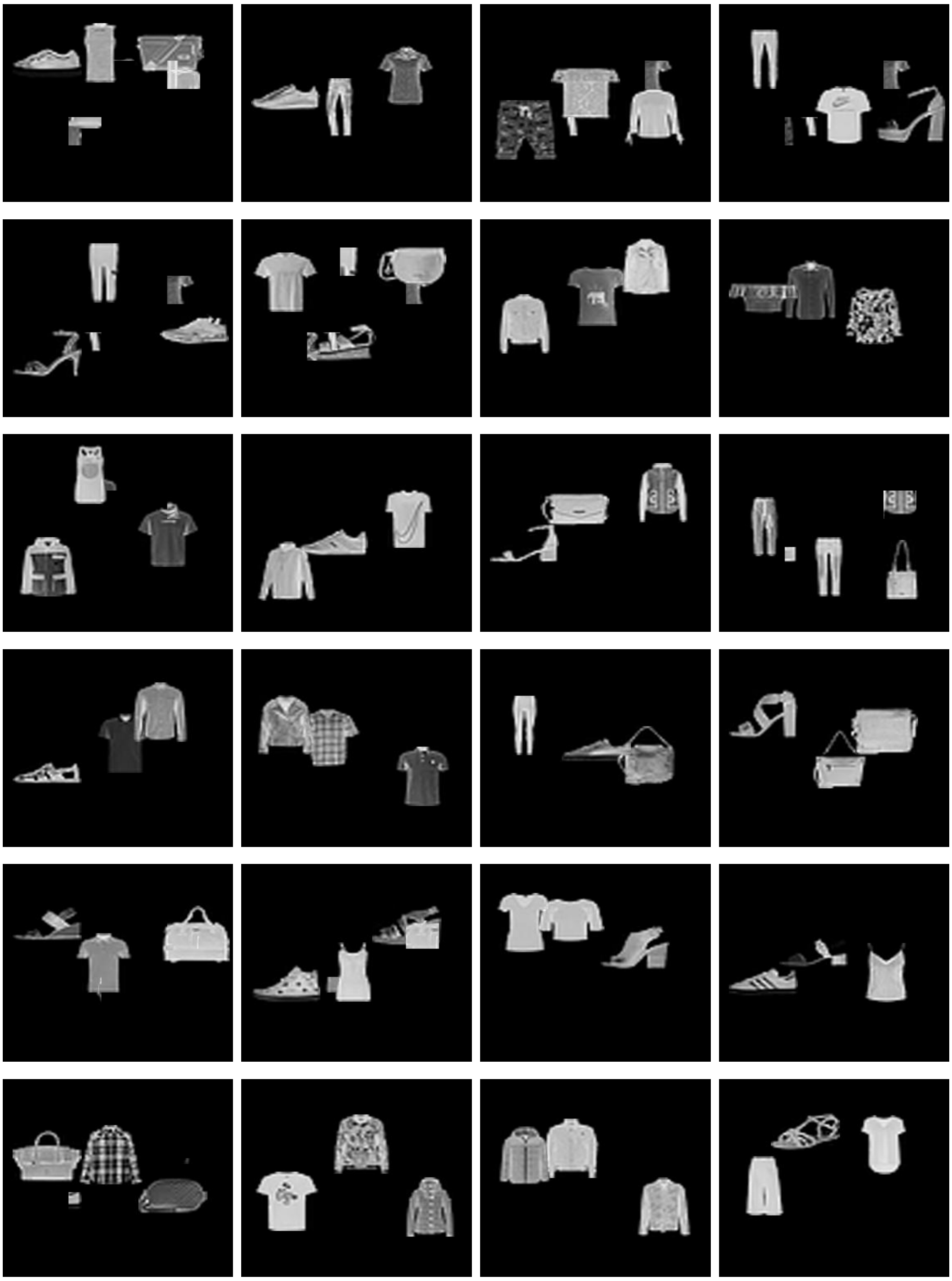


Figure 3.5: Randomly selected images from Fashion-ZSD dataset.

CHAPTER 4

ZERO-SHOT OBJECT DETECTION IN THE WILD

4.1 Overview

Zero-shot learning represents approaches for handling image classification [206, 207, 208], object detection [55, 56] [1], instance segmentation [209], image captioning [2] and other problems [210, 211] with objects belonging to classes that are not seen in the training set. Over the past decade, studies on zero-shot learning have mainly focused on the image classification task [206, 207, 208, 99, 100, 85, 88], but there has been an increasing attraction to the ZSD in recent years [55, 56, 115, 120, 118, 117]. The main reason for the interest in ZSD is that the object detection problem needs more labor-intense data labeling compared to the classification problem. In such a case, ZSD is an alternative way to reduce data annotation costs for object detection by adapting models to detect classes not included in the training set.

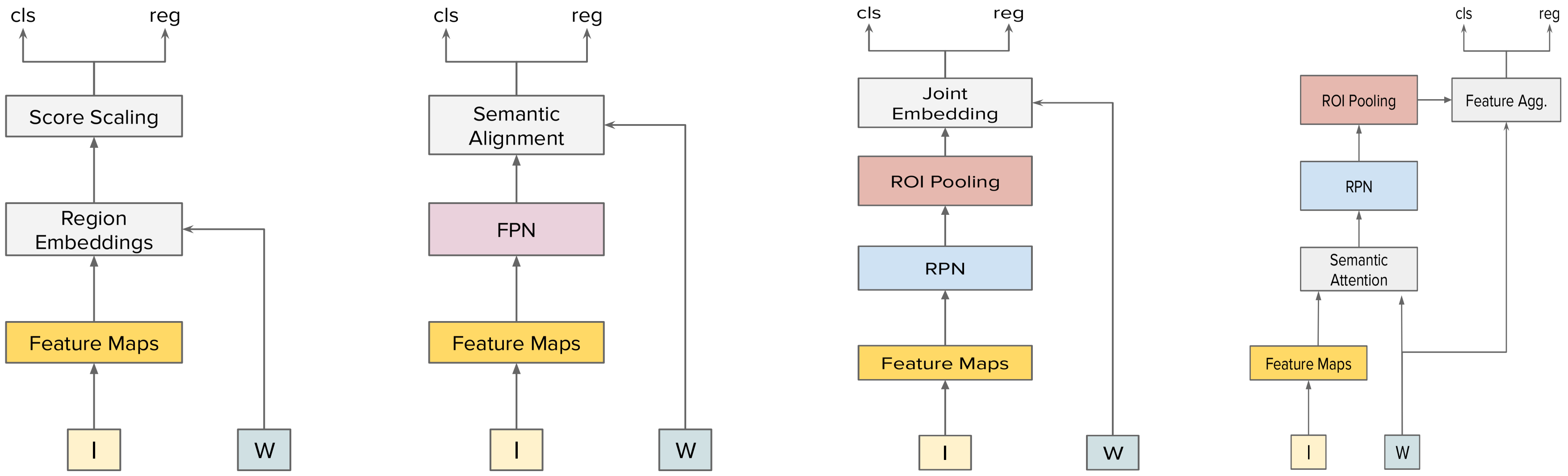
Supervised object detection, which aims to both localize and classify objects, has gained great momentum with high detection performance of deep learning models [15, 16, 18, 17, 20, 21, 23, 24, 212, 40]. The recent ZSD approaches are also based on these supervised single-stage [15, 21] or two-stage [17, 212] object detection models. In these ZSD approaches, the aim is to handle the class similarities using semantic side-information, which is usually obtained from class embeddings.

Among the ZSD studies, Rahman *et al.* [55] defines polarity loss to better align the compatibility between visual and semantic domains. In this way, visual relations between semantic class embeddings are encoded more accurately. Yan *et al.* [56]

---

[1] In addition, our works mentioned in Chapter 3 and Chapter 6 are among these approaches.

[2] Our work mentioned in Chapter 6.



(a) SimEmb 6 (b) PL [55] (c) C-ZSD [56] (d) Ours

Figure 4.1: Comparison of most relevant ZSD models in the literature.

proposes a semantic-guided contrastive learning method to transform word embeddings into visually meaningful form. Our work mentioned in Chapter 3 proposes a single-stage ZSD model to generate visual embedding vectors from the object detector in addition to bounding box coordinates and classification scores. Their model performs object detection using both these visual embeddings and convex combinations over classification scores. Gupta *et al.* [117] proposes a transformed semantic space that uses both visual and semantic spaces as complementary to each other, and combines multi-space scores to obtain predictions. Our work mentioned in Chapter 6 balances the prediction bias for the seen and unseen classes by learning the scaling parameters and uncertainty from training data. Li *et al.* [115] uses natural language explanations instead of class embeddings of classes as semantic information. In this context, word-word and word-visual attentions are used jointly to generate object candidates.

Contemporary object detection training schemes sample image regions with low ground-truth annotation overlap to collect negative samples, and ZSD training schemes adopt the same training strategies. In such a ZSD model training, if there are instances of unlabeled classes in the training images, these class instances might be selected and learned as background regions, and thus it is difficult to generate candidate boxes at the inference stage for these unlabeled class instances. In order to avoid this problem, to the best of our knowledge, ZSD approaches remove images consisting of instances of unseen classes from the training sets [55, 56]. However, this situation is against the nature of the ZSD problem in two aspects and creates a dilemma: (i) images consisting of selected unseen classes should be known and discarded during the training time, (ii) the image level annotations of the selected unseen classes are already known during training.

To overcome this issue, we propose an approach for object detection in unseen classes without the above-mentioned curated training set in the ZSD problem, and our key idea is to customize the feature maps of input images for each target class. In this way, customized feature maps do not model unseen instances as background regions in training and generate candidate proposals for unseen classes in inference even if instances of these classes appear in the training set. At this point, we use semantic side-information (*e.g.* class embeddings) to do semantic attention on the visual feature maps and obtain class-specific features for the region proposal network (RPN) [17] and subsequent layers in our two-stage object detection model.

Figure 4.1 shows an overview of the previously proposed one-stage (Fig. 4.1a and 4.1b) [55] and two-stage (Fig. 4.1c) [56] ZSD models. In this figure, gray boxes represent model-specific components. I and W represent the input image and class embeddings, respectively. According to this figure, previously proposed ZSD approaches use feature maps regardless of the class, without any filtering or attention. Thus, if instances of unlabeled unseen classes are available in the training set, it is possible for these classes to be modeled as background regions by object detection models. To this end, we propose a two-stage ZSD model (Fig. 4.1d) that performs semantic attention on backbone features, unlike other approaches. Therefore, our model can more effectively be trained even on images containing (unknown) samples of unseen classes. In our model, attention is performed on the pre-RPN feature maps by using positive and randomly selected negative class embeddings. While the proposed ZSD method also needs to use negative object proposals for training, these proposals are generated over the feature maps which use these class embeddings. In the detection head, visual and semantic features are concatenated, and these joint features are fed into the separate classification and regression heads. Extensive experiments on MS-COCO [39] and Pascal VOC [79] datasets show that the performances of ZSD methods are greatly decreased when instances of unlabelled unseen classes are present during training. Again, the same experiments show that our proposed method obtains successful results even if the instances of unseen classes exist in the training set.

In summary, our work provides the following contributions to the ZSD problem:

- We examine the background modeling problem for ZSD and propose a first attempt to handle it, to the best of our knowledge.
- We show experimentally that the performance scores of several state-of-the-art models greatly drop when trained in a more realistic setting where training images contain unseen classes.
- We propose a semantic attention mechanism and use customizable feature maps according to the input side information.
- According to the experiments, our approach greatly diminishes the effect of unseen class samples in the background and obtains state-of-the-art results on

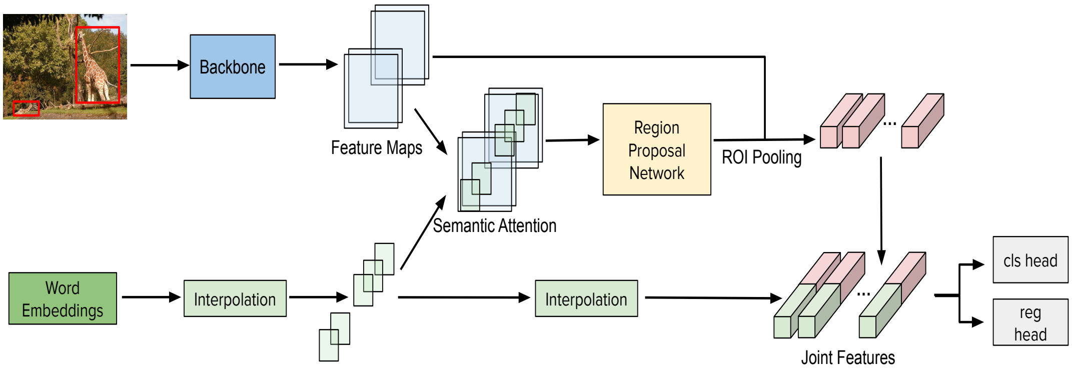


Figure 4.2: An overview of the proposed network architecture.

benchmark datasets when images containing unseen classes are not discarded from the training set.

4.2 Method

Our ZSD network (Figure 6.2) is based on the ubiquitous two-stage Faster R-CNN model [17]. In our architecture, input images $x \in R^{h,w,3}$ are fed into a backbone network and feature maps $f \in R^{\frac{h}{s}x\frac{w}{s}xC}$ are obtained. Here, h and w depict the height and the width of the input image, C is the number of channels, and s denotes how much the spatial size of the image is reduced. These visual feature maps are attended with semantic embeddings ϕ (*e.g.* word2vec [50], Glove [51]) via depth-wise cross correlation [213]. After that, a set of anchor boxes on each feature map point are placed. Objectness scores and regression offsets of all anchor boxes are learned during the training with binary cross-entropy and smooth-L_1 losses, respectively.

As in a typical zero-shot learning setting, we have semantic embeddings of seen class names and a large amount of annotated visual data for seen classes, but have no visual or semantic information for unseen classes during training. Meanwhile, we have semantic word embedding vectors for all classes at inference time.

In the following subsections, we first describe the semantic attention module enabling us to generate category-specific region proposals as a key component of our approach (Section 4.2.1). Then, a general view of our two-stage detection network is presented

(Section 4.2.2). Finally, how inference works on instances never seen either visually or semantically during the training stage is explained (Section 4.2.3).

4.2.1 Semantic Attention

We use a similar paradigm with the Attention-RPN module [189], which is proposed for the few-shot object detection problem, to modify the input image feature maps. In this context, we convolve the input image feature maps with fixed kernels rather than learning the kernel weights through training. Different from the Attention-RPN approach which uses support images as kernels to generate class-specific proposals for few-shot classes, we use class embeddings to modify feature maps for background modelling:

$$f^{'}(i,j) = f(i,j) \odot \phi(c_k) \tag{4.1}$$

where, $\phi(c_k)$ and $f(i,j)$ denote the semantic word embedding vector of the k-th object class that appears in the input image, and feature map points in f, respectively. $\odot$ represents the Hadamard product, and $f^{'}(i,j)$ is used to generate category-specific region proposals. The proposed semantic attention mechanism needs to use negative object proposals for training. Hence, we perform semantic attention on the pre-RPN feature maps by using positive and randomly selected negative class embeddings. In order to apply semantic attention, we need to align the dimensions of visual feature map channels and class embeddings, so we apply nearest-neighbor interpolation on the class embeddings to align vector dimensions.

The proposed network establishes relations between semantic and visual information of classes and learns to generate region proposals specific to the object categories by guiding RPN losses with the content of the modified feature map and ground truth bounding box of the related object.

4.2.2 Zero-shot Detector

RPN generates a fixed number of proposals P for each object category. Proposals are denoted as axis-aligned rectangles with top-left and bottom-right coordinates. As rectangles have varying sizes, ROI Pooling operation [16] is used to produce same-sized ROIs $R_p \in R^{NxKxKxC}$ from modified feature maps. Here, N, K, and C denote the number of proposals, the spatial dimension of each ROI, and the number of feature channels, respectively. After obtaining ROI features R_{P_i} for each proposal, we use a residual block to further enrich instance features and apply Global Average Pooling (GAP) [214] operation to transform feature tensors to vectors. As residual block double instance feature channels, we double the word vector dimensionalities with a second nn-interpolation. Finally, visual feature vectors and corresponding word vectors are concatenated:

$$J = R_{P_i} \parallel \phi(c_k) \tag{4.2}$$

Here, i and k denotes the i-th candidate proposal and k-th positive or random negative training class. Moreover, J represents the joint features to use in classification and regression heads as used in [189]. We use separate fully connected networks with two layers and ReLU activation function to make fine adjustment on proposed rectangles and classify them as either foreground or background. Again, smooth-l_1 loss and cross entropy loss are used for regression and classification in this stage like in [17]. Finally, the total loss function can be represented as follows:

$$L_{total} = L^{rpn}_{BCE} + L^{rpn}_{L_1} + L^{roi}_{BCE} + L^{roi}_{L_1} \tag{4.3}$$

Here, L^{rpn}_{BCE} and L^{roi}_{BCE} represent the binary cross entropy classification losses of region proposal network and detection head, respectively. $L^{rpn}_{L_1}$ and $L^{roi}_{L_1}$ denote least absolute deviation (L_1) losses of region proposal network and detection head, respectively.

4.2.3 Inference for Unseen Classes

Class embeddings of visually similar classes are also closer to each other [55]. This semantic similarity property of class embeddings enables to generate proposals for classes that have not seen visually during training time. In the literature, ZSD methods use benchmark MS-COCO and Pascal VOC dataset splits which manually remove images in the training set that consisting of unseen class instances. Otherwise, these models tend to learn unseen class object candidates as background regions.

In our proposed method, the RPN uses features that are convolved with the different semantic word vectors during training and inference time. Thanks to the this property of our method, we do not have to remove images consist of novel object instances from the training set and create a more realistic zero-shot setting. At inference time, we generate separate object proposals for each novel class by using word vectors one at a time. These proposal rectangles are either kept or eliminated according to the classification scores and kept ones are thoroughly regressed. The proposed method generates separate proposals for each unseen class, so there is no need to make a multi-class classification.

4.3 Experiments

In this section, we present the details of the proposed method and comparisons with existing ZSD methods. In this context, we share the details of the semantic attention, class embeddings, and ZSD model details in Section 5.3.2. Then, we compare and discuss the proposed method with existing ZSD models in Section 4.3.2.

4.3.1 Implementation Details

We train our model with a batch size of 4 and learning rate of 0.001. We use ImageNet pre-trained ResNet-101 network ($s = 16$, $C = 1024$) as the backbone network and pretrained 300-dimensional GloVe embedding vectors[51] as side information for all of the our experiments. The maximum number of object proposals for RPN is selected as $N = 2000$ for training and $N = 40$ for inference. As emphasized earlier, we do not

Table 4.1: Experimental results on MS-COCO (65/15), Pascal VOC (16/4) and Pascal VOC (17/3) datasets.

Method	MS-COCO (65/15)	VOC (16/4)	VOC (17/3)
HRE 3	4.12 (↓12.8)	15.89 (↓54.20)	-
PL [55]	4.59 (↓12.40)	-	21.21 (↓42.50)
SimEmb 6	5.45 (↓15.78)	17.23 (↓57.47)	-
SA	**12.92**	**27.62**	**37.14**

remove images containing unseen class objects from the training sets.

4.3.2 Main Results

We compare the proposed Sematic Attention (SA) method on MS-COCO and Pascal VOC benchmark datasets using ZSD splits that are proposed in HRE and PL methods. We use HRE, PL, and SimEmb studies as baselines for comparisons. Since the PL does not share the trained class embeddings for the Pascal VOC split (16/4), in which there are 4 selected unseen classes, we identify 3 Pascal VOC classes in common with the MS-COCO unseen classes (*i.e.* train, cat, airplane) as unseen test classes and prepare Pascal VOC experiments with 17/3 settings. We chose all of the state-of-the-art models for which we could find the source code as baselines. We are unable to make comparisons with the remaining ones as they report in the unrealistic setting presuming the availability of unseen class annotations. We use 0.5 IoU threshold value as in other methods.

We share the obtained results in Table 4.1. Accordingly, the ZSD methods experience major mAP drops in our more realistic evaluation scenario. While the SimEmb method achieves a score of 15.78 mAP for the unseen classes in the MS-COCO dataset, this score becomes 5.45 with a great loss of 10.33 mAP when instances of unseen classes are not removed from the training set. Similarly, on the Pascal VOC dataset, the HRE method looses 38.31 mAP, while the PL method lost 21.29 mAP in newly proposed splits. In contrast, our proposed method achieves state-of-the-art results with a large margin on both MS-COCO and Pascal VOC datasets. We share some visual results in

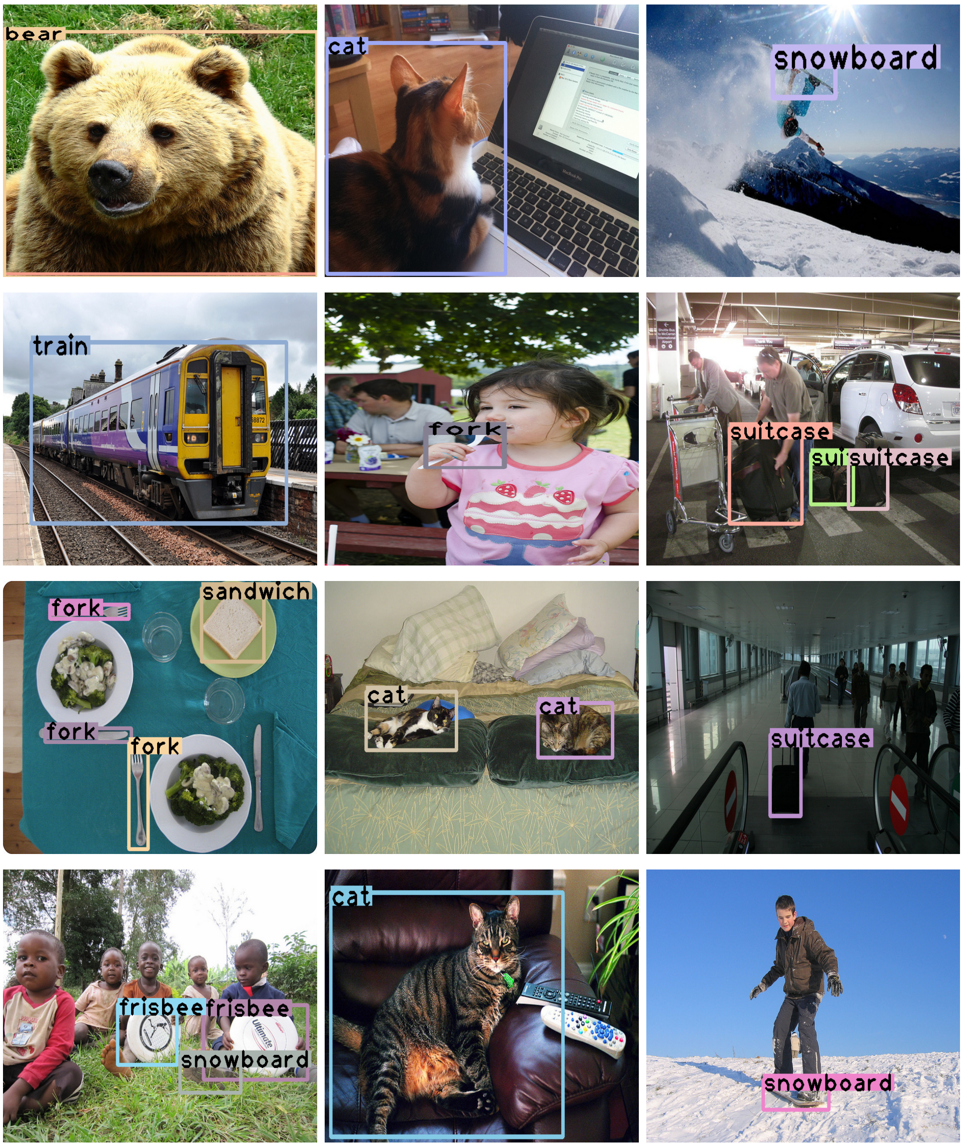


Figure 4.3: Some visual ZSD results on MS-COCO dataset. (Best viewed in color.)

Figure 4.3. Accordingly, the proposed method obtains results for unseen classes even though they are background in the training set.

We also show the effect of the maximum number of candidate proposals for each class on the mAP in Figure 7.3.2. It is observed that mAP values increase in a non-monotonic way as the number of proposals increases until $N = 120$. Our ZSD model reaches its highest mAP value with **13.04** when $N = 100$. This ablation is not applicable to the other methods since they do not generate class-specific object candidates.

Finally, we repeat the MS-COCO experiments by discarding images containing in-

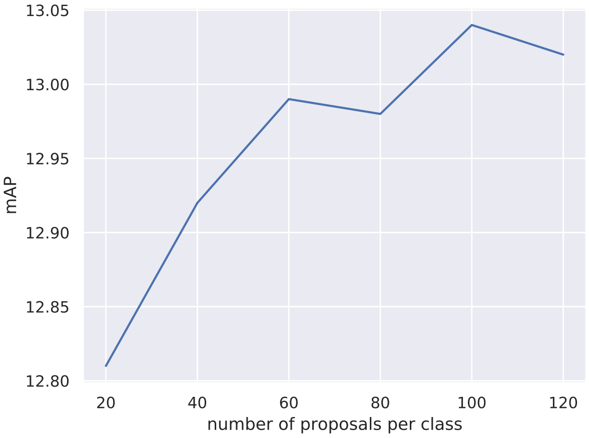


Figure 4.4: mAP values for different maximum number of object proposals for RPN.

stances of unseen classes from the training set as in other approaches. In these experiments, our method obtains 13.76 mAP for the unseen classes. This shows that the proposed method does not model unseen classes as backgorund even though they exist in the training set.

4.4 Chapter Summary

One of the most important shortcomings of the current ZSD methods is that they know that there are instances of unseen classes in the training set and the images containing these instances are eliminated in order to produce successful ZSD models. With our proposed method, pre-RPN features become class-specific, so that even if the images containing these instances are in the training set, it can also generate and classify candidate proposals for unseen classes at inference time. To the best of our knowledge, this paper is the first attempt in the ZSD literature that draws attention to the background modeling problem. Experimental results show that the proposed method obtains promising results on benchmark MS-COCO and Pascal VOC datasets.

CHAPTER 5

LEARNING VISUALLY CONSISTENT LABEL EMBEDDINGS FOR ZERO-SHOT LEARNING

5.1 Overview

With the surge of deep learning models, there is a high demand of large-scale datasets for training classification models over a large number of classes. However, annotating such large-scale data is both highly costly and labor-intensive. Zero-shot learning (ZSL) emerges as a promising alternative in this regard. ZSL is a form of learning to handle classification when the labelled training data is available for only some of the classes (called *seen classes*, *i.e.* training classes), and, not for the others (called *unseen classes*, *i.e.* test classes). The basic philosophy of this technique is transferring knowledge from seen to unseen classes by utilizing prior information from various sources such as textual descriptions of classes (*e.g.* [215, 216, 2]), embeddings of class names (*e.g.* [52, 53, 54]) or attribute-based class specifications (*e.g.* [215, 217, 88, 218, 52, 54, 219]). Overall, the performance of a ZSL method heavily depends on the prior information as it is the primary factor determining the limits of cross-class knowledge sharing and transfer.

In this study, we aim to increase the success of label-embedding based ZSL models by incorporating visually meaningful word vectors for class embeddings. More specifically, the word embeddings of class names used in label embedding techniques are typically derived from textual information in previous work [52, 53, 54]. These word vectors tend to capture only semantic relations, ignoring the visual resemblances between the corresponding visual concepts. We argue that this may cause a considerable loss of information for ZSL for object recognition. Instead, we propose to ground

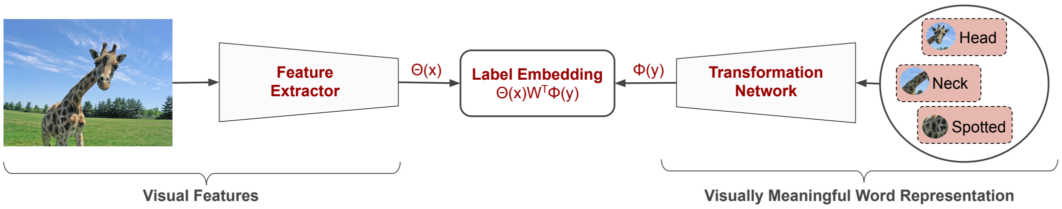


Figure 5.1: We propose a zero-shot learning approach based on visually meaningful word vectors and label embedding.

label embeddings on visually meaningful word vectors proposed by [219], which aims to transform word embeddings such that each class name and the corresponding combination of attribute names attain a high degree of similarity. Unlike [219], however, instead of relying directly on the attribute-to-class associations in the transformed word embedding space, we construct our final ZSL model using the image-to-class associations measured by a label-embedding classifier.

In this work, we explore this idea and leverage visually meaningful word vectors as auxiliary data in label embedding to cover the bottlenecks of the both techniques. For this purpose, we learn visually more consistent word vectors and embedding space in an end-to-end manner by defining a joint loss function. This approach is illustrated in Figure 5.1. While using label embeddings, our approach utilizes the word representations transformed to a visually more consistent space. At test time, our zero-shot learning approach allows assigning novel images to unseen classes, purely based on class names.

To sum up, our main contribution in this work is utilizing the visually meaningful class name embeddings obtained by learning to associate corresponding attribute combinations and class names, and use them within a label embedding framework, without requiring human-annotated attribute-class relations for the unseen classes.

In our experiments, we evaluate the proposed idea on two ZSL benchmark datasets, namely Animals with Attributes (AwA) [220] and aPascal-aYahoo (aPaY) [217] datasets. We use word vectors which are obtained from GloVe method [51] to represent textual data. We also use CNN-M2K features [218] to represent visual features and learn attribute based classifiers. Our experimental results show that our method yields encouraging improvements in recognition accuracy on these benchmark datasets.

5.2 Method

In this paper, we build an end-to-end framework based on visually consistent word vectors and label embeddings. Basically, our method learns a transformation network that maps word vectors to an embedding space more suitable for zero-shot learning. For this purpose, we propose a framework to jointly learn the word embedding transformation and the label embedding models in an end-to-end manner.

In the rest of the section, we present the details of our approach. We first give a brief summary of the approach of [219], which we utilize for learning visually meaningful word vectors, and then describe how we use these vectors as the side information in the label embedding model.

5.2.1 Visually Meaningful Vector Space Word Vectors

The introduction of distributed word vector representations, such as Word2Vec [50] or GloVe [51], has been a step forward in semantic word representations, since these representations tend to capture the semantic nuances and relations between words more accurately. Based on their success, these vector space representations have witnessed a great attention in ZSL techniques and a large variety of other applications ranging from document retrieval to question answering. Nevertheless, in computer vision problems, semantic similarities at the word level may not be enough to model all the variances of the visual categories. For example, semantically similar words, such *"wolf"* and *"bear"* are not particularly close in visual domain, whereas visually consistent words such as *"mole"* and *"mouse"* can be far apart in semantic word domain. In order to account for such differences, [219] propose to learn a transformation on the word vectors that allows ZSL by comparing the pooled embeddings of attribute names and class names. Below we provide only a brief summary of the *image-based training* formulation of this approach, a more through explanation can be found in [219].

In this formulation, the similarity between the class y_i of an image x_i and the set of associated attributes recognized in it should be higher than its similarity when another class embedding is used (y_j):

$$s(x_i, y_i) \geq s(x_i, y_j) + \Delta(y_i, y_j), \qquad \forall y_j \neq y_i \tag{5.1}$$

where Δ is a margin function, indicates pairwise discrepancy value for each given training classes. In this inequality, $s(x, y)$ represents a compatibility function that measures the relevance between a pair of class and a set of posterior-probability weighted attributes, formulated through a multilayer perceptron network. It also corresponds to a mapping that allows the transformation of word vectors from semantic to visually meaningful space. This approach is formalized as a constrained optimization problem:

$$\min_{\Phi,\xi} \lambda ||\Phi||_2^2 + \sum_{i=1}^{N} \sum_{y_j \neq y_i} \xi_{ij}$$
$$s(x_i, y_i) \geq s(x_i, y_j) + \Delta(y_i, y_j) - \xi_{ij} \quad \forall y_j \neq y_i, \forall i$$

where λ is the regularization weight. Here, we learn a transformation matrix, which then we will call as Φ. We refer to this transformation network as A2CN.

5.2.2 Label Embedding

In order to use the visually consistent word vectors with visual data for ZSL, we prepare an embedding method:

$$f(x, y; W) = \Theta(x)^T W \Phi(y) \tag{5.2}$$

where visual descriptors are denoted by $\Theta(x)$ and textual side information is denoted by $\Phi(y)$. Moreover, W matrix encodes textual and visual data to assign unseen test classes to correct class labels. This matrix is designed as a dense layer in the multilayer perceptron network. In the Eq. 5.2, $\Phi(y)$ side information are feeded from A2CN model outputs. Cross-entropy loss is used to learn a proper embedding space. Softmax classifier is also applied to normalise network predictions so that results can be interpreted as probabilities. Finally, we use the following joint loss function to learn transformation and embedding networks in an end-to-end manner. The final learning formulation, therefore, takes the following form:

$$\min_{\Phi,\xi,W} \sum_{i=1}^{N} \sum_{y_j \neq y_i} \xi_{ij} - \sum_{i=1}^{N} \log \frac{\exp \Theta(x_i)^T W \Phi(y_i)}{\sum_{y' \neq y_i} \exp \Theta(x_i)^T W \Phi(y')}$$

$$s(x_i, y_i) \geq s(x_i, y_j) + \Delta(y_i, y_j) - \xi_{ij} \quad \forall y_j \neq y_i, \forall i$$

where ℓ_2 regularization is additionally applied to the parameters W and Φ, but omitted from the equation for brevity.

5.3 Experiments

In this section, we present the details of our experiments. First, we give a brief information about the datasets, and then explain initial word embeddings and visual features. Then, we give the details of our experiments, where we compare the proposed approach with its unsupervised and supervised counterparts.

5.3.1 Datasets

To evaluate the proposed approach, we use two benchmark ZSL datasets, namely Animals with Attributes (AwA)[220] and aPascal-aYahoo (aPaY)[217]. The AwA dataset consists of images with 50 animal classes, 40 of which are training and 10 of which are test. 85 per-class attributes are defined on these classes. aPaY dataset consists of images from two different sources. Training part is obtained from Pascal VOC 2008 [221] dataset, containing 20 classes. The test part is collected using Yahoo search engine and it contains 12 classes; totaling up to 32 completely different classes overall. Images in aPaY dataset are annotated with 64 binary per-image attributes. We follow the same experimental setup as in [219] for AwA and aPaY experiments.

5.3.2 Implementation Details

Initially, we use 300-dimensional word embedding vectors which are obtained from GloVe method as described in A2CN method [219] for fair comparison. Following the

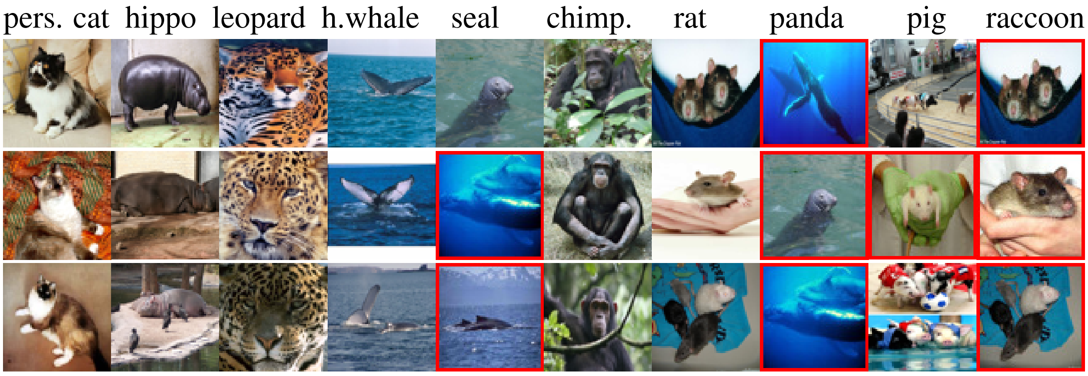


Figure 5.2: Top-3 highest scoring images using PBT method in the AwA dataset.

A2CN method, we obtain word vectors for each class and attribute names. If attribute or class names consist of multiple words, word vectors are obtained for each word, then the average of these vectors is used.

For AwA and aPaY datasets, we utilize the CNN-M2K features [218], where images are resized to 256x256 and mean image subtraction is applied. Outputs of the last hidden layer are extracted for image representation, as also described in [219].

We define our method as a three layer feed-forward network. We use 2-fold cross-validation to determine optimal number of hidden units. Adam optimizer [222] is used for stochastic optimization and the learning rate value is set to 1e-4. The proposed joint loss function is used to learn transformation and embedding networks in an end-to-end manner.

5.3.3 Experimental Results

In our experiments, we first evaluate our method using two different training methods, PBT and IBT, that are proposed by [219] to handle ZSL problem. PBT stands for *(Predicate-based Training)* and IBT stands for *(Image-based Training)*. We measure the performance using the normalized per-class accuracy and the results are shown Table 5.1. According to the obtained results, it seems that our method provides a noticeable progress using PBT training method, where there is a 2.9% accuracy increase in AwA and 4.9% increase in aPaY datasets. For the IBT method, some improvements are also observed in the aPaY dataset, whereas the recognition performance slightly

Table 5.1: Zero-shot classification performance of proposed method on AwA and aPaY datasets.

Test	Method	AwA	aPaY
PBT	A2CN[219]	60.7	29.4
	Our Method	**63.6**	**34.3**
IBT	A2CN[219]	**69.9**	38.2
	Our Method	68.6	**40.8**

degrades on AwA dataset. This may be due to the fact that the parameters learned during the cross-validation may not produce the best results for the test classes. We also believe that PBT method is more important than IBT, because it only uses predicate matrix to learn meaningful word vectors, so it is a more generalizable method with less training data.

We also compare our approach with various unsupervised and supervised counterparts presented in the literature. The results are shown on Table 5.2. Here, supervised methods require additional information about test classes such as class-attribute relations. Unsupervised ZSL methods do not require any human supervision about the unseen test classes. When we review the results on Table 5.2, we observe that our method obtains higher classification performance on aPaY dataset, compared to its unsupervised counterparts. On AwA dataset, it outperforms its unsupervised counterparts, except for A2CN[219] method.

Our ZSL method also produces comparable results to some of the supervised counterparts. Another interesting direction to note is that, while high accuracies can potentially be obtained using the recently proposed data generation models [93, 223, 87, 224], these works are orthogonal to proposed method, and, in principle, these techniques can be used in combination with the ZSL model proposed in this work. We plan to investigate this line of research in future work.

Finally, Figure 5.2 illustrates qualitative examples of the results of the our approach. In this figure, we show the top-3 scoring images produced by our proposed method on AwA dataset. The misclassified images are marked with red. According to this

Table 5.2: Comparison of the related ZSL literature.

Test Supervision	Method	AwA	aPaY
unsupervised	DeViSE [84]	44.5	25.5
	ConSE [86]	46.1	22.0
	Text2Visual [2, 225]	55.3	30.2
	SynC [81]	57.5	-
	ALE [53]	58.8	33.3
	LatEm [52]	62.9	-
	CAAP [54]	67.5	37.0
	A2CN [219]	**69.9**	38.2
	Our Method	68.6	**40.8**
supervised	DAP [220]	54.0	28.5
	ENS [45]	57.4	31.7
	HAT [218]	63.1	38.3
	ALE-attr [53]	66.7	-
	SSE-INT [226]	71.5	44.2
	SynC-attr [81]	76.3	-

illustration, misclassifications tend to occur across visually similar classes, as expected.

5.4 Chapter Summary

The performance of the zero-shot learning approaches depend on the shared prior information between training and unseen test classes; therefore, it is very critical that the prior information is accurate, consistent and comprehensive. In this work, we have aimed to improve zero-shot recognition by using visually meaningful word vectors within the label embedding framework. The experimental results show the effectiveness of the proposed approach.

CHAPTER 6

CAPTION GENERATION ON SCENES WITH SEEN AND UNSEEN OBJECT CATEGORIES

6.1 Overview

The problem of generating a concise textual summary of a given image, known as *image captioning*, is one of the most challenging problems that require joint vision and lingual modeling. With ever-increasing recognition rates in object detection models, pioneered by [15, 14, 16, 18, 19, 17, 20, 21, 22, 23, 24], there has been a recent interest in generating visually grounded captions via constructing detection-driven captioning models, *e.g.* [135, 147, 227, 57]. However, the success of such approaches is inherently limited by the set of classes spanned by the detector training set, which is typically too small to construct a visually comprehensive model. Therefore, such models are prone to synthesizing irrelevant captions in realistic, uncontrolled settings where input images may contain instances of classes unseen during training.

In the context of image classification, *zero-shot learning* (ZSL) has emerged as a promising alternative towards overcoming the practical limits in collecting labeled image datasets and constructing image classifiers with very large object vocabularies. In a similar manner, *zero-shot image captioning* (ZSIC), aims to develop methods towards overcoming the data collection bottleneck in image captioning. However, we observe that there is no prior work irectly tailored to study captioning in a truly zero-shot setting, except the preliminary conference version of this paper to the best of our knowledge: recent works on ZSIC [57, 58] study the ZSIC problem only in the language domain, presuming the availability of a pre-trained fully-supervised object detector covering all object classes of interest. We refer to these methods as *partial*

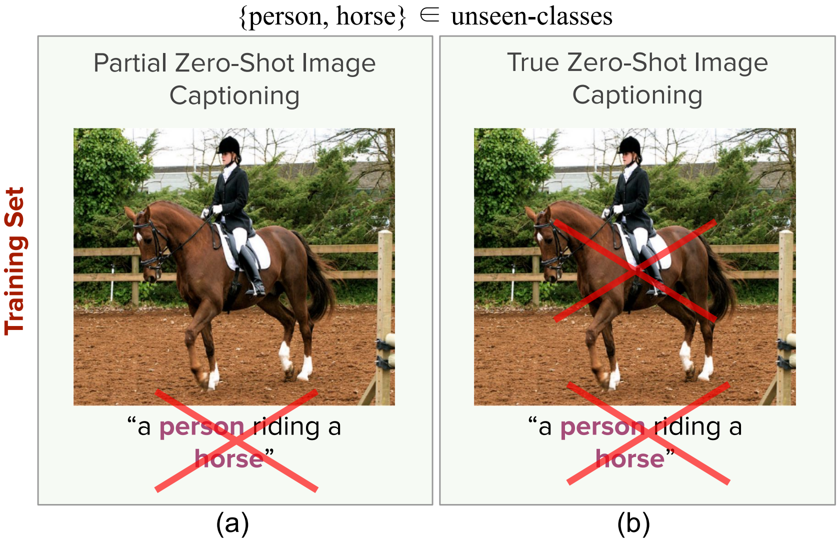

{person, horse} ∈ unseen-classes
Training Set
Partial Zero-Shot Image Captioning
True Zero-Shot Image Captioning
"a person riding a horse"
"a person riding a horse"
(a)
(b)

Figure 6.1: (a) *Partial zero-shot image captioning* problem, (b) *True zero-shot image captioning* problem.

zero-shot image captioning.

Following these observations, we propose the problem of *true zero-shot captioning*, where test images contain instances of unseen object categories with no supervised visual or textual examples, in addition to the seen categories. We believe that this change constitutes a more direct problem definition towards (i) developing semantically scalable captioning methods, and, (ii) evaluating captioning approaches in a realistic setting where not all object classes have training examples. The difference between the partial versus true ZSIC problems is illustrated in Figure 6.1.

To tackle the true ZSIC problem, we propose an approach that consists of a novel generalized zero-shot detection (GZSD) model, which aims to generate detections in scenes with both seen and unseen class instances, and a template-based [57] caption generator. A high-level summary of our ZSIC approach can be found in Figure 6.2. In order to address the GZSD problem, we propose a scaling scheme and incorporate *uncertainty calibration* [125] to make seen and unseen class scores comparable. We also show out that using class-to-class similarities obtained over word embeddings [50] as *class embeddings* improves the GZSD results, compared to using class name embeddings directly. On the MS-COCO dataset [39], we present a detailed evaluation

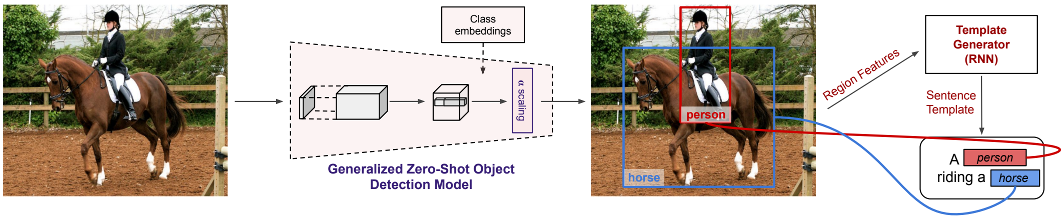


Figure 6.2: Our zero-shot captioning framework.

of both GZSD and ZSIC models. For a more accurate evaluation of the ZSIC results, we propose a new evaluation metric called V(isual)-METEOR, which adapts and improves the widely used METEOR metric for ZSIC evaluation purposes.

A preliminary version of this work aims to make a number conceptual, technical and experimental contributions in image captioning, which can be summarized as follows:

- We define a new paradigm for generating captions of unseen classes.
- We propose a novel ZSD approach that incorporates a probability scaling scheme for the generalized zero-shot object detection (GZSD) problem.
- We evaluate several caption evaluation metrics and discuss their suitability for the zero-shot image captioning scenario.

In addition to provide more detailed related work discussions and method explanations, this paper extends the conference version by:

- introducing uncertainty calibration loss for class confidence calibration,
- evaluating the impact of various model decisions and score calibration,
- introducing a comparison to the recent GZSD methods on the benchmark MS-COCO dataset,
- quantitatively demonstrating the advantage of using class-to-class similarities as the class embeddings,
- and analyzing the GZSD failure patterns, which are all directly relevant for the captioning quality.

The journal version also proposes the V-METEOR metric, and uses the new metric for a more detailed analysis of the ZSIC model.

6.2 Method

In this section, we first explain our main ZSD model component, and its GZSD extensions. We then explain how we build the ZSIC model. Finally, we discuss the evaluation difficulties and define the V-METEOR metric.

6.2.1 Main zero-shot detection model

In ZSD, the goal is to learn a detection model over the examples given for the seen classes (Y_s) such that the detector can recognize and localize the bounding boxes of the unseen classes Y_u. For this purpose, we adapt the YOLO [14] architecture to the ZSD problem.

In the original YOLO approach, the loss function consists of three components: (i) the localization loss, which measures the error between ground truth locations and predicted bounding boxes, (ii) the objectness loss, and (iii) the recognition loss, over a prediction grid of size $S \times S$. Following our prior conference work, we adapt the YOLO model to the ZSD problem by replacing per-cell class probability predictions with *cell embeddings* and re-defining the prediction function as a compatibility estimator between the cell and class embeddings:

$$f(x, c, i) = \frac{\Omega(x, i)^T \Psi(c)}{\parallel \Omega(x, i) \parallel \parallel \Psi(c) \parallel}. \tag{6.1}$$

Here, $f(x, c, i)$ is the prediction score corresponding to the class c and cell i, for image x, $\Psi(c)$ represents the c-th class embedding, and $\Omega(x, i)$ denotes the predicted cell embedding as shown in Figure 6.3. According to this figure, at each cell, the network is trained to produce box coordinate predictions (denoted by b_x, b_y, b_h, b_w in the figure), objectness scores (denoted by s in the figure) and a cell embedding to be used for zero-shot recognition. The resulting model, therefore, allows making detection predictions for samples of novel classes purely based on their class embeddings.

Class embeddings. In principle, one can use attributes or word embeddings of class

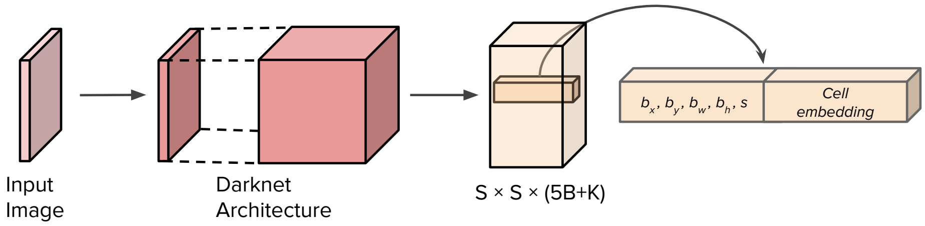


Figure 6.3: Summary of the proposed GZSD method.

names directly as class embeddings, *e.g.* Chapter 3. Attributes can provide powerful visual descriptions of classes, however, they tend to be domain-specific and typically difficult to define for a large variety of object classes, as needed in ZSIC. Word embeddings of class names are much easier to collect, however, they typically contain indirect information about the visual characteristics of classes, and therefore, known to provide significantly weaker prior knowledge for visual recognition [53].

To use the word embeddings more effectively, we propose to define class embeddings in terms of class-to-class similarities computed over word embeddings: we define the class c embedding in terms of the similarity with each seen class $\bar{c}$:

$$\Psi(c) = \left[\varphi(c)^T \varphi(\bar{c}) + 1\right]_{\bar{c} \in Y_s} \tag{6.2}$$

where $\varphi(c)$ denotes the c-th class name's word embedding. Since semantic relations across classes tend to correlate with their visual characteristics, this embedding can provide a valuable implicit visual description defined through a series of inter-class similarities. The ZSL method, therefore, can make predictions based collectively on these similarity values. We empirically demonstrate the advantage of this scheme in Section 6.3.

6.2.2 Generalized zero-shot detection extensions

There can be a significant bias towards the seen classes as the GZSD model is trained to predict seen class instances. We use the following two extensions to reduce this bias.

Alpha scaling. In this technique, we aim to reduce the bias towards the training classes

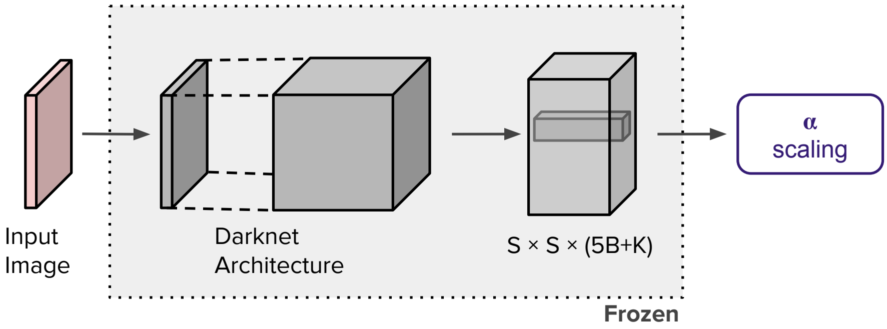


Figure 6.4: α scaling factor learning process.

by making the unseen and seen class scores more comparable through a score scaling scheme. For this purpose, we introduce the α coefficient for the unseen test classes, and redefine $f(x, c, i)$ as follows:

$$f(x,c,i) = \begin{cases} \alpha \frac{\Omega(x,i)^T \Psi(c)}{\|\Omega(x,i)\| \|\Psi(c)\|}, & \text{if } c \in Y_u \\ \\ \frac{\Omega(x,i)^T \Psi(c)}{\|\Omega(x,i)\| \|\Psi(c)\|}, & \text{otherwise} \end{cases} \tag{6.3}$$

To make the α estimation practical, we want to avoid requiring additional training examples. For this reason, we first train the ZSD model over all training classes without α. We then designate a subset of seen classes as *unseen-imitation* classes. To obtain unseen-like confidence scores for these classes, we temporarily set all entries corresponding to unseen-imitation classes in Eq. 6.2 to zeros and treat unseen-imitation classes as unseen classes in Eq. 6.3. These modifications allow us to obtain classification scores as if the model was trained without using the samples of unseen-imitation classes. We then train α only, keeping the rest of the network frozen, as shown in Figure 6.4.

Overall, the proposed α coefficient estimation scheme leverages the special structure of our class embeddings to efficiently approximate the unseen class scores. While the approximation can possibly be coarse, we experimentally show in Section 6.3 that the proposed scheme is effective for learning the α coefficient, at a negligible extra training cost.

Uncertainty calibration. The second unbiasing technique that we explore is *uncer-*

tainty calibration, adapted from the zero-shot classification approach of Liu *et al.* [125]. The idea is to minimize the uncertainty over unseen class predictions during training, based on the observation that a prediction model learned over seen class samples tends to yield lower confidence scores for unseen classes, resulting in misdetections.

The uncertainty in confidence scores is quantified via entropy over unseen class probabilities. We adapt the uncertainty calibration loss ℓ_h to our ZSD model as a loss over per-cell predictions:

$$\ell_h(x) = -\sum_{i=0}^{S^2} \mathbb{1}_{\text{obj}}^{i} \sum_{c \in Y_u} p_u(c|x,i) \log p_u(c|x,i) \tag{6.4}$$

Here, $p_u(\cdot)$ corresponds to $f(x,c,i)$-driven unseen class likelihoods:

$$p_u(c|x,i) = \frac{\exp(f(x,c,i)/\tau)}{\sum_{c' \in Y_u} \exp(f(x,c',i)/\tau)} \tag{6.5}$$

where τ denotes the softmax temperature coefficient. τ is empirically determined as in Liu *et al.* [125]. The loss encourages more confident unseen class score estimates, as less ambiguous prediction results in smaller entropy values. In order to adapt the uncertainty calibration to the detection model, we first train the ZSD model over all training classes as in the alpha scaling optimization process. We also use the same designated unseen-imitation subset as unseen classes. In the second training stage, we temporarily set all entries corresponding to unseen-imitation classes to zeros and then fine-tune the whole model without freezing any layers, unlike alpha scaling coefficient learning.

6.2.3 Zero-shot captioning model

Our goal is the construction of an image captioning model that can accurately summarize scenes potentially with seen and unseen class instances. For this purpose, we opt to use a template-based captioning method which provides the sentence templates with visual word slots to be filled based on the outputs of an object detection model.

We adapt the slotted sentence template generation model of *Neural Baby Talk* (NBT) [57]. The NBT method generates sentence templates which consist of the empty word slots by using a recurrent neural network. To obtain a content-based attention mechanism over the grounding regions, NBT embraces *pointer networks* [228]. The NBT model is trained by optimizing the model parameters ω such that the log-likelihood of each ground-truth caption q conditioned on the corresponding image x is maximized:

$$\omega^* = \text{arg} \max_{\omega} \sum_{(x,q)} \log p(q|x;\omega). \tag{6.6}$$

Here, the conditional caption likelihood $p(q|x;\omega)$ of $|q|$ words is measured auto-regressively, using a recurrent network:

$$p(q|x;\omega) = \prod_{t=1}^{|q|} p(q_t|q_{1:t-1}, x;\omega). \tag{6.7}$$

The NBT method additionally incorporates a latent variable r_t to represent the specific image region, so the probability of a word q_t is modeled as follows:

$$p(q_t|q_{1:t-1}, x;\omega) = p(q_t|r_t, q_{1:t-1}, x;\omega)p(r_t|q_{1:t-1}, x;\omega). \tag{6.8}$$

The NBT defines two word types for q_t, corresponding to *textual* and *visual* words. Textual words are not directly related to any image region or specific visual object instance, therefore the model provides only dummy grounding for them. The template generation network uses the object detection outputs to fill empty visual word slots, where we utilize the outputs of our GZSD model.

We train both the GZSD model and the sentence template generation component of NBT over examples containing only the seen class instance annotations, as required by the *true ZSIC* protocol. At test time, we use the GZSD outputs over all classes as inputs to the NBT sentence generator.

6.2.4 Measuring zero-shot captioning quality

Partial zero-shot image captioning approaches use existing captioning metrics, such as METEOR [59], SPICE [155] and F1 score, for evaluation purposes. While these generic textual similarity based metrics provide useful information about the quality of captioning results, they do not explicitly handle the problem of capturing visual content within the generated sentence. Therefore, such metrics can possibly be heavily influenced by structural and syntactic similarities across generated and ground-truth sentences. Exceptionally, F1 score differs in this regard by completely ignoring the sentence structure and measuring only the coverage of (unseen) class names within captions. However, F1 score fails to measure the overall quality or accuracy of the generated sentences, which is also clearly important.

We observe that, based on our experiments in Section 6.3, the explicit handling of visual and non-visual content in the evaluation of sentences is particularly necessary for *true zero-shot image captioning*. In this setting, the problem of generating sentences that summarize the visual content accurately, including visual entities that are completely unseen during training, is fundamentally challenging, especially in comparison to partial ZSIC with fully-supervised visual recognition models. Therefore, we propose a new captioning evaluation metric as a step towards formalizing better metrics for true ZSIC.

We develop our metric based on METEOR, which is known to be a simple yet effective metric that yields a strong correlation with human judgment [229]. The original METEOR metric is defined by the following formula:

$$\text{METEOR} = F_{\text{mean}}(1 - p) \tag{6.9}$$

where F_{mean} aims to capture correctness in terms of unigram precision and recall values and p is a penalty term for evaluating the overall sentence compatibility. More specifically, F_{mean} is given by:

$$F_{\text{mean}} = \frac{10PR}{R + 9P} \tag{6.10}$$

where P and R are the unigram precision and unigram recall values, respectively. These are calculated as:

$$\begin{aligned} P &= \frac{m}{w_t} && (6.11) \\ R &= \frac{m}{w_r} && (6.12) \end{aligned}$$

where m is the number of unigrams in both reference and generated captions, w_t is the number of unigrams in the candidate caption and w_r is the number of unigrams in the reference caption. The p penalty term checks how well textual chunks match between a pair of reference and generated captions, using the following definition:

$$p = 0.5 \left(\frac{s_c}{u_m} \right)^3 \tag{6.13}$$

where s_c is number of maximally long matching subsequences, and u_m is number of mapped unigrams.

We extend the METEOR metric by defining two separate F_{mean} metrics for the visual and non-visual entities. For this purpose, we compute F^v_{mean} and F^n_{mean}, similar to Eq. 6.10, separately over only visual words and only non-visual words, respectively. We, then, define the proposed metric V-METEOR based on their harmonic mean, as follows:

$$\text{V-METEOR} = \frac{2F^v_{\text{mean}}F^n_{\text{mean}}}{F^v_{\text{mean}} + F^n_{\text{mean}}}(1-p) \tag{6.14}$$

In this manner, the proposed V-METEOR metric explicitly measures the joint visual or non-visual accuracy of a sentence, through the harmonic mean of the F^v_{mean} and F^n_{mean} terms. It also incorporates the overall sentence similarity by keeping the penalty term (p) as in METEOR.

To be able to measure per-class captioning quality, which is particularly valuable in the ZSIC context, we separately compute V-METEOR for each class. In the calculation of the V-METEOR score of a sentence for a class, the words corresponding to the class name are considered as the visual words, and the words that are not corresponding to any one of the class names are considered as non-visual words. The overall V-METEOR score is obtained by averaging per-class scores.

Finally, we additionally define the following two variations for separately measuring the visual and non-visual quality of the generated sentences, respectively:

$$\text{V-METEOR}_{\text{vis}} = F^v_{\text{mean}}(1-p) \tag{6.15}$$

$$\text{V-METEOR}_{\text{nvis}} = F^n_{\text{mean}}(1-p) \tag{6.16}$$

We use V-METEOR$_{\text{vis}}$ and V-METEOR$_{\text{nvis}}$ to gain additional insights.

6.3 Experiments

In this section, we explain our experimental setup, present the GZSD and ZSIC results, discuss the V-METEOR evaluations, and provide additional analyses.

6.3.1 Experimental setup

ZSD and (partial) ZSIC works use different splits of the MS-COCO dataset for historical reasons. To make our results comparable to related works, we use the same splits as in the related works, separately for GZSD and ZSIC as explained below.

GZSD evaluation. We use MS-COCO [39] dataset in our experiments. In our main GZSD experiments, we use the same dataset splits and settings as in the recent works [117, 118, 119, 120, 121, 56, 55, 114] and our ZSD method in Chapter 3, where 15 of 80 MS-COCO classes are used as unseen classes. There also exist different ZSD methods (*e.g*. SB [112] and DSES [112]), but they use only 48/17 seen-unseen class distribution or do not share GZSD results with 65/15, so we do not report any comparisons with these methods.

ZSIC evaluation. For the ZSIC approach, we compare the proposed approach with selected *upper-bound* methods from [147, 148, 58, 149, 57]. We again use the same dataset splits and settings as in these works, where 8 of 80 MS-COCO classes are used as the unseen classes.

Word embeddings. For the GZSD model, we use 300-dimensional word2vec [230] class name embeddings. For the names containing more than one word, *e.g*. *tennis racket*, we take the average of the per-word embeddings. We use 300-dimensional GloVe vector embeddings [51] in the template generation component of the ZSIC, following the NBT approach [57].

6.3.2 Generalized zero-shot object detection

In this section, we report and discuss experimental results for the GZSD model. We train the model for 160 epochs with a learning rate of 0.001, and a batch size of 32. Once the model is trained, we select 8 out of 65 seen classes as *unseen-imitation* classes for alpha scaling optimization and uncertainty calibration purposes, and continue training for 10 more epochs.

Main results. We present the experimental results in Table 6.1. The upper part of the table presents results of the two-stage object detection techniques, and the lower part presents the single-stage techniques and our approach, which we call *SimEmb*. In the lower part, **SimEmb-base**, which represents the model without score calibration, obtains 28.54% mAP on seen classes, 12.45% mAP on unseen classes and 17.34 harmonic mean (HM). **SimEmb**, which represents the version with learned α scaling coefficient, obtains 28.91% mAP on seen classes, 15.78% mAP on unseen classes and 20.41% HM. Finally, **SimEmb*** represents an upper-bound reference model, where alpha scaling coefficient is empirically tuned on the test set to maximize the HM score by evaluating for a range of α values. This upper-bound model obtains 28.87% mAP on seen classes, 16.00% mAP on unseen classes, and 20.59 HM value.

From the results, we first observe that our single-stage approach improves the state-of-the-art among single-stage GZSD models. We also observe that SimEmb performs similar to or better than many two-stage GZSD models, with the only exception being the very recently published two-stage approach ContrastZSD [56]. Second, the improvements obtained by SimEmb show that alpha scaling coefficient is crucial for obtaining higher accuracy on unseen class detections and alpha scaling does not disrupt the seen class performance. Finally, the comparison between SimEmb and the SimEmb* upper-bound shows that the proposed alpha scaling learning scheme is effective as it yields results comparable to directly tuning α on the test set.

We also observe that the proposed model achieves results comparable to those of two-stage approaches. While single-stage and two-stage detectors are built on very different design principles and trade-offs, the overall competitiveness is noteworthy since the work on other low-shot detection problems show that two-stage models typically yield higher AP scores [62].

Qualitative detection results using the proposed SimEmb model can be found in Figure 6.5.

Correctness of α estimation. We present the evaluation results as a function of α in Figure 6.6. We observe that the best empirical α coefficient value (in HM) among the tested ones is 1.4. The proposed α estimator, which in contrast uses only training examples, results in $\alpha = 1.28$, which is both value-wise and performance score-wise close to the optimal choice.

Table 6.1: mAP results on MS-COCO dataset with GZSD (65/15) settings.

Category	Method	seen	unseen	HM
two-stage	MS-Zero [117]	**42.40**	12.90	19.79
	MS-Zero++ [117]	35.00	14.50	19.78
	DPIF-S [118]	32.72	13.95	19.56
	DPIF-M [118]	29.33	16.36	21.00
	BLC [119]	36.00	13.10	19.20
	VL-SZSD [120]	39.45	13.18	19.76
	FNG [121]	38.10	13.90	20.40
	ContrastZSD [56]	40.20	**16.50**	**24.20**
single-stage	TL [114]	28.79	14.05	18.89
	PL [55]	**34.07**	12.40	18.18
	Chapter 3	28.40	12.80	17.65
	SimEmb-base	28.54	12.45	17.34
	SimEmb	28.91	15.78	20.41
	SimEmb*	28.87	**16.00**	**20.59**

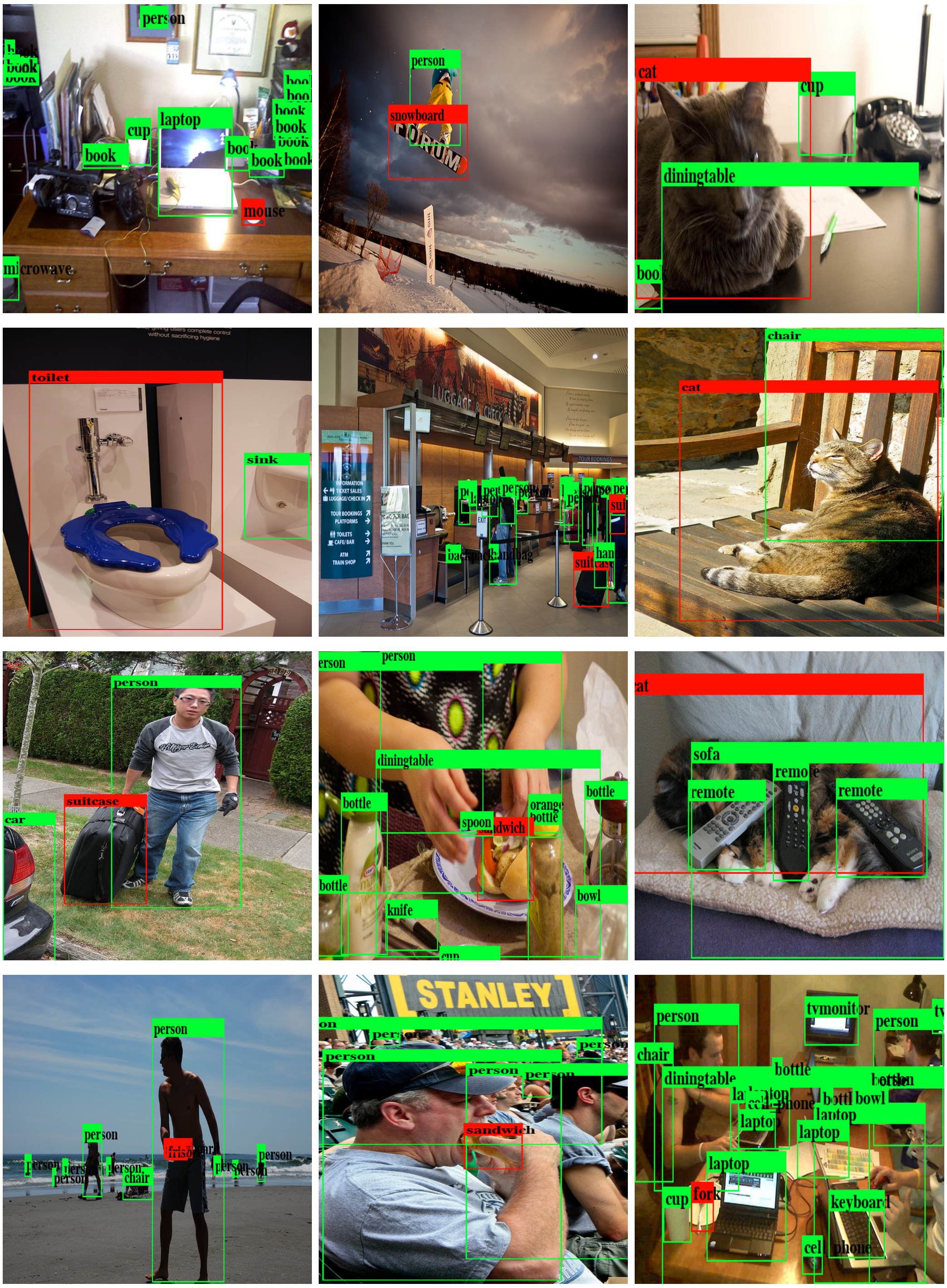


Figure 6.5: GZSD results on scenes containing various **seen** and **unseen** class instances.

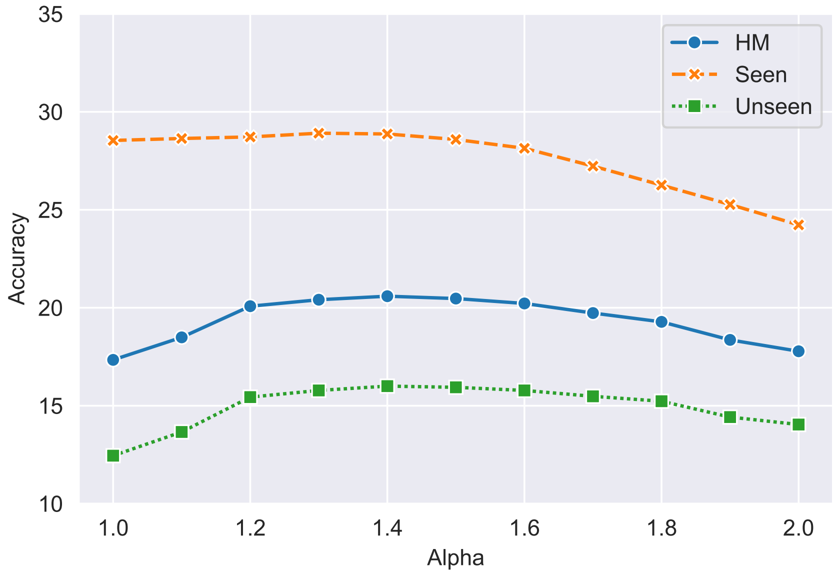


Figure 6.6: The accuracy values of the proposed method in the GZSD test splits of MS-COCO according to different alpha scaling factors.

Alpha scaling versus uncertainty calibration. As an alternative to alpha scaling for GZSD, we evaluate the uncertainty calibration technique, as explained in Section 6.2.2. We present the results in Table 6.2, with the following combinations from top to the bottom: base model, uncertainty calibration (*uc-calib*) only, alpha scaling only, and their combination. We observe that uncertainty calibration alone performs poorly probably due to the difficulty of correcting class bias purely based on fine-tuning. Our alpha scaling technique yields a much better result in terms of HM score, with an improvement from 17.34 to 20.41. The combination of the two techniques slightly improves the HM score to 20.46. This proves that the alpha scaling scheme is effective in comparison to a state-of-the-art calibration technique. For the sake of simplicity, we keep using only alpha scaling in our following experiments.

GZSD results on ZSIC splits. In our experiments presented so far, we have used the 65/15 COCO split. In our ZSIC experiments, however, we need to use the alternative 72/8 split of [147] to make comparisons to the related work. Therefore, here we report the results of our GZSD model on the 72/8 split. We train the model using the same hyper-parameters as before. We select 8 out of 72 seen classes as unseen-imitation

Table 6.2: mAP results on MS-COCO dataset in the 65/15 GZSD setting.

α-scaling	uc-calib	seen	unseen	HM
		28.54	12.45	17.34
	✓	28.60	11.15	16.04
✓		**28.91**	15.78	20.41
✓	✓	28.85	**15.85**	**20.46**

classes for alpha scaling optimization.

We evaluate the detection model under the ZSD and GZSD scenarios. For the ZSD experiments, we use the MS-COCO validation images consisting of unseen class instances. For the GZSD experiments, we use the whole MS-COCO val5k split. We present the results on Table 6.3. Here, the first row represents the experimental results where we only use images belonging to the unseen classes and unseen class embeddings, the remaining rows represent the GZSD results where we use all class embeddings on the MS-COCO val5k split. In the ZSD case, we observe an unseen class mAP of 31.4%. In the GZSD case, we observe a much lower 0.3% mAP without alpha scaling, and 0.7 HM. Alpha scaling improves the unseen class mAP to 7.3% and the HM score to 10.6. We note that prior works on GZSD do not use this ZSIC (72/8) split, therefore, we do not report any comparisons to the state-of-the-art in this split. We also note that our primary interest in GZSD is to build a strong method to serve as a crucial component of ZSIC, therefore, these results highlight one of the major difficulties in building accurate captioning models in the realistic ZSIC setting.

6.3.3 Zero-Shot image captioning

For the ZSIC experiments, we use the same experimental setup described in [57], and exclude the image-sentence pairs containing unseen class instances during training. We consider the partial ZSIC approaches proposed in [147, 148, 58, 149, 57] as upper-bound baselines for our true ZSIC setting. We also define and evaluate a baseline method based on NBT, where we train the NBT captioning model based solely on

Table 6.3: Our results on ZSD and GZSD (72/8).

Exp. Type	Test	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	U-mAP(%)	S-mAP(%)	HM
ZSD	U	5.2	53.3	35.1	23.9	44.2	36.4	9.1	43.7	31.4	-	-
GZSD w/o α	S+U	0	0	2.7	0	0	0	0	0	0.3	27.4	0.7
GZSD	S+U	0.8	21.4	4.9	1.2	4.5	0.7	9.1	15.8	7.3	19.2	10.6

the training classes without integrating our GZSD model. We refer to this model as *NBT-baseline*.

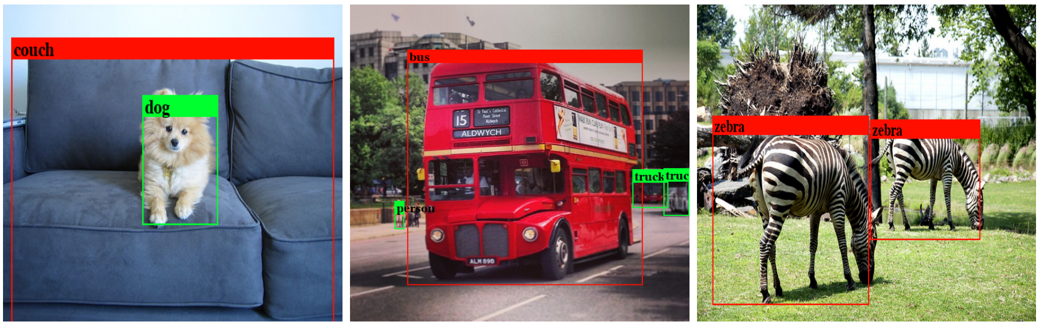
To establish a fair comparison, we follow the practices of the NBT [57] approach. We evaluate the ZSIC model on the selected validation subset of the MS-COCO caption dataset. To obtain per-class evaluation scores, we use the F1 metric [147], where a visual class is considered as relevant in an image if that class name appears in any one of the human generated reference captions for that image, and irrelevant otherwise. Similarly, on a test image, a model-generated caption is considered as correct for a visual class if the generated caption includes (excludes) the corresponding word for that relevant (irrelevant) class. The per-class F1 score is then defined as the ratio of correctly captioned test images. We additionally use the well-established METEOR [59] and SPICE [155] metrics, in addition to averaging the per-class F1 scores (referred to as *Avg. F1*). We separately discuss the evaluation results in terms of the proposed V-METEOR metric in the next section.

We present the results in Table 6.4. First, we observe that the proposed approach greatly outperforms the NBT-baseline with clear improvements in terms of Avg. F1 (0 to 29.8), METEOR (18.2 to 21.9) and SPICE (12.7 to 14.2) scores. This shows the value of explicitly handling the GZSD task as part of the captioning process. In comparison to the upper-bound partial-ZSIC captioning approaches, which involve supervised visual training in both seen and unseen classes, our approach yields comparable results in terms of METEOR and SPICE metrics. In particular, we observe that the ZSIC model yields better results compared to the DCC [147] and NOC [148] methods. This is most probably due to the fact that our sentence template generation method provides accurate locations for visual words, enabling the generation of more natural and visually grounded captions. We observe relatively lower scores for the ZSIC model, compared to the remaining supervised models.

Noticeably, the performance gap between true ZSIC and (visually) supervised partial ZSIC is larger in terms of the Avg. F1 metric. This is mostly an expected result as the F1 metric directly measures the ability to incorporate visual classes during captioning, akin to a visual recognition metric. Here, supervised methods are known to perform much better than the state-of-the-art ZSL models in most cases, which turns out to

Table 6.4: Zero-shot captioning results with comparison to captioning models involving visually fully-supervised models.

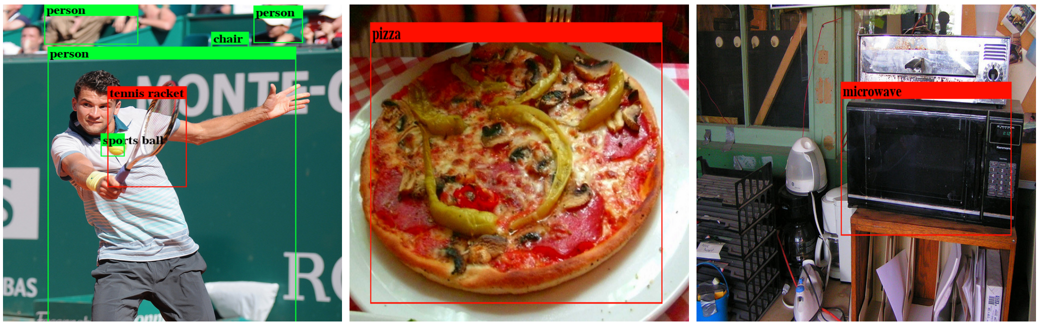
Method	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg. F1	METEOR	SPICE
True zero-shot captioning											
NBT-baseline	0	0	0	0	0	0	0	0	0	18.2	12.7
Our method	2.4	75.2	26.6	24.6	29.8	3.6	0.6	75.4	29.8	21.9	14.2
Partial zero-shot captioning (upper-bounds)											
DCC [147]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8	21.0	14.4
NOC [148]	17.8	68.8	25.6	24.5	69.3	68.1	39.9	89.0	49.1	21.4	-
C-LSTM [58]	29.7	74.4	38.8	27.8	68.2	70.3	44.5	91.4	55.7	23.0	-
Base+T4 [149]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0	23.3	15.9
NBT+G [57]	14.0	74.5	42.8	63.7	74.4	19.0	44.5	92.0	53.2	23.9	16.6
DNOC [150]	33.0	77.0	54.0	46.6	75.8	33.0	59.5	84.6	57.9	21.6	-



A small white *dog* sitting on a **couch**.

A red **bus** is driving down the street.

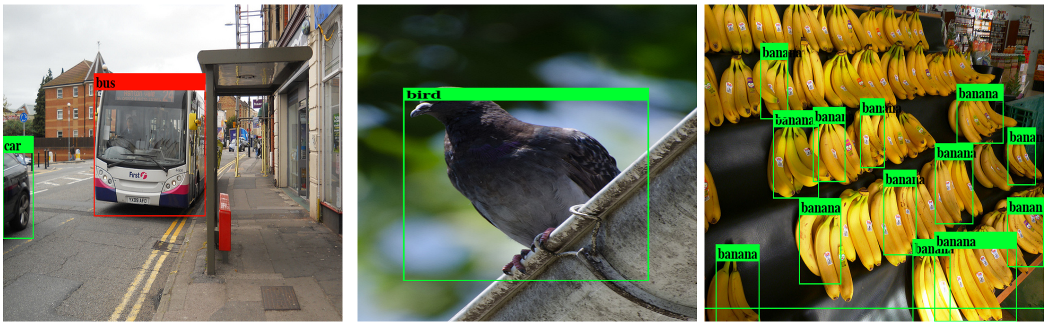
A couple of **zebra** standing in a field.



A *tennis player* is about to hit a **tennis racket**.

A white plate topped with a piece of **pizza**.

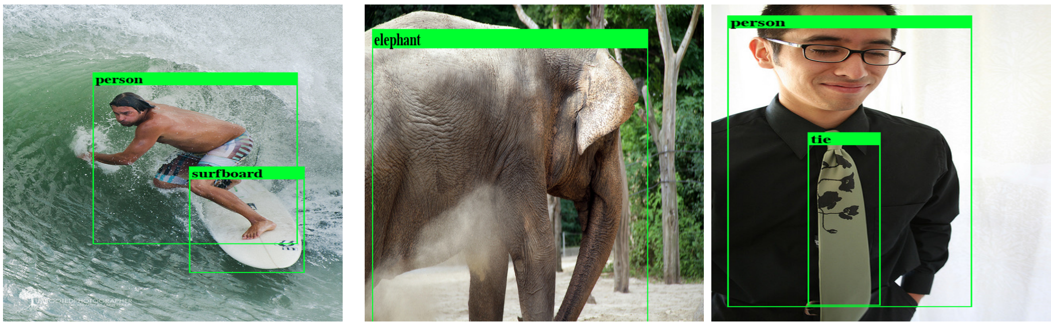
A kitchen with a **m.wave** and a counter.



A **bus** is parked on the side of the street.

A *bird* sitting on top of a metal pole.

A bunch of *banana* that are on a table.



A *man* riding a wave on top of a *surfboard*.

A large *elephant* standing next to a tree.

A *man* in a suit and *tie* standing in a room.

Figure 6.7: Image captioning results on images with *seen* and **unseen** class instances.

Table 6.5: V-METEOR comparison results.

Method	V-METEOR$_{\text{vis}}$	V-METEOR$_{\text{nvis}}$	V-METEOR
NBT-Baseline	0.0	20.50	0.0
Our Method	12.63	22.26	13.19

also be the case in captioning.

For qualitative examination, we present visual output examples in Figure 6.7, along with the corresponding GZSD detection results. It can be observed that the ZSIC model is able to generate semantically sound captions in a variety of challenging scenes involving both seen and unseen class instances.

6.3.4 V-METEOR experiments

We now evaluate the baseline and proposed models using the V-METEOR metric. We present the overall average V-METEOR scores in Table 6.5. In this table, V-METEOR$_{\text{vis}}$ represents a sub-metric that only includes results for visual words, and V-METEOR$_{\text{nvis}}$ represents an another sub-metric that only includes non-visual words. These summary results show that the proposed approach greatly improves the visual captioning score from 0.0 to 12.63 and also increases the non-visual V-METEOR scores from 20.50 to 22.26. The final V-METEOR score improves from 0.0 to 13.19. These results show that the integration of an (accurate) GZSD can not only help with visual coverage of the captioning results but also improve the non-visual parts of the generated captions thanks to the better visual information from the detector to the language model. In these results, we also observe the main advantage of the proposed V-METEOR metric by being able to separately discuss the visual and non-visual quality of the generated captions.

To better understand the captioning results, we present per-class V-METEOR scores for the unseen classes in Figure 6.8, where **visual-bs** represents the visual meteor scores of the NBT-Baseline, **non-visual-bs** represents the non-visual meteor scores of the NBT-Baseline and **hm-bs** represents the V-METEOR scores of the NBT-Baseline

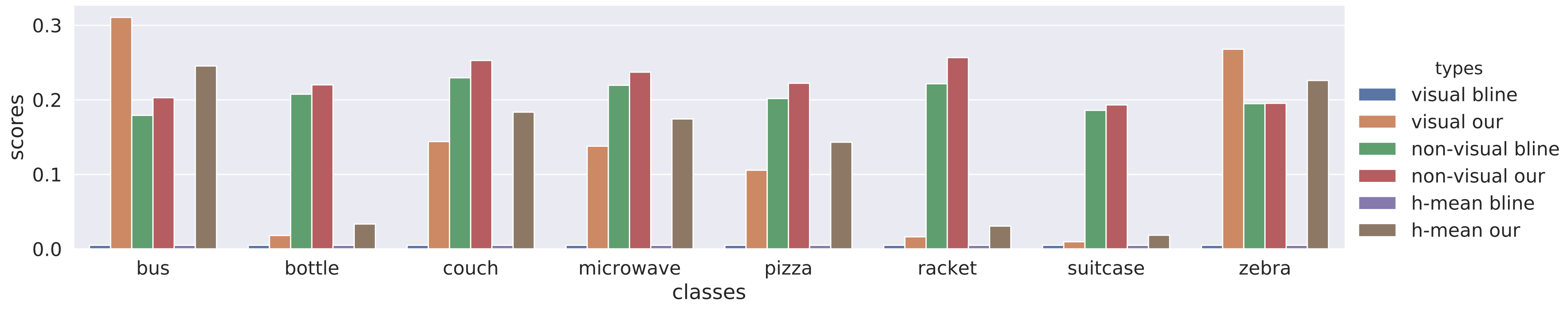
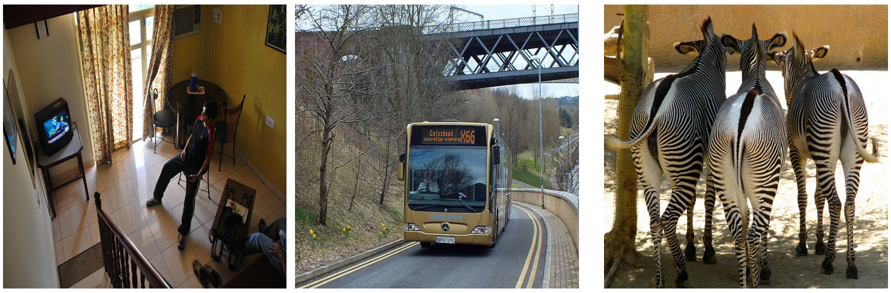


Figure 6.8: V-METEOR results of each unseen classes.



♦: A couple of **people** that are in a room.
★: A **person** sitting in a **couch** in a room.

♦: A yellow and black **train** traveling down the road.
★: A yellow and black **bus** driving down a road.

♦: A couple of **elephants** standing next to each other.
★: A couple of **zebra** standing next to each other.

Figure 6.9: Image captioning results of NBT-baseline and our methods.

method. Similarly, **visual, non-visual** and **hm** bars correspond to our method. In these results, we again observe both the most significant improvements are in V-METEOR$_{\text{vis}}$ scores with still noticeable improvements in non-visual scores. The complementary qualitative captioning comparisons presented in Figure 6.9, where ♦ represents the NBT-baseline results, ★ represents the results of the proposed method, and **bold** type words represent visual words from detectors, supports these quantitative observations: in the *person* and *bus* examples, the whole sentence changes and improves with the correction in visual details. In the *bus* and *zebra* examples, we observe that the NBT-baseline method produces coarsely plausible sentences, however, with incorrect visual coverage due to confusions across visually similar classes.

6.3.5 Additional analyses

In this section, we present a quantitative analysis on the error patterns and an ablative study on the importance of proposed similarity embeddings in GZSD.

6.3.5.1 Diagnosing errors

The experimental results show that GZSD plays a central role in achieving accurate captioning results. Therefore, it is potentially valuable to understand the typical detection errors of our GZSD model, towards building better GZSD and ZSIC approaches.

For this purpose, we embrace the detector analysis approach by Hoiem *et al.* [231], which is originally proposed for analyzing false positives in supervised detectors. The original analysis approach defines semantic categories for the PASCAL VOC dataset. To utilize this technique in the GZSD setting, we use the MS-COCO superclasses, namely *vehicle, outdoor, animal, accessory, sports, kitchen, food, furniture, electronic, appliance* and *indoor*, as defined in [39]. Following [231], we additionally define a separate singleton superclass for the *person* class, as it contains a greatly larger number of instances and its overall distinct visual characteristics.

The following four misdetection categories are examined for each superclass: (i) localization errors, corresponding to detections considered as false positive due to poor localization, (ii) confusion with background, counting false positive detections located in the background, (iii) class confusion within superclass members, and (iv) class confusion across superclasses. The corresponding error distributions are shown in Figure 6.10.

The obtained error distribution results show that the false positives are mainly occurred due to the within superclass confusions for the *vehicle, animal, accessory, sports, kitchen* and *food* superclasses. The dominant misdetection type for the *furniture, appliance* and *indoor* superclasses is confusion with other classes. In contrast, most *person* misdetections correspond to localization errors. Finally, we observe that most problematic detections for *outdoor* and *electronic* superclasses correspond to background detections. Overall, these results show that there is no single error pattern dominating the GZSD outputs, and errors vary greatly across the classes.

6.3.5.2 Impact of using similarity embeddings

One of the advantages of using the proposed class-to-class similarity vectors is that each dimension of the embedding explicitly corresponds to a class relevance value. We additionally utilize its structure in the design of our alpha scaling training scheme. To better understand the value of the proposed class embeddings for GZSD, we present a direct comparison between using the proposed class embeddings versus the original class name word embeddings.

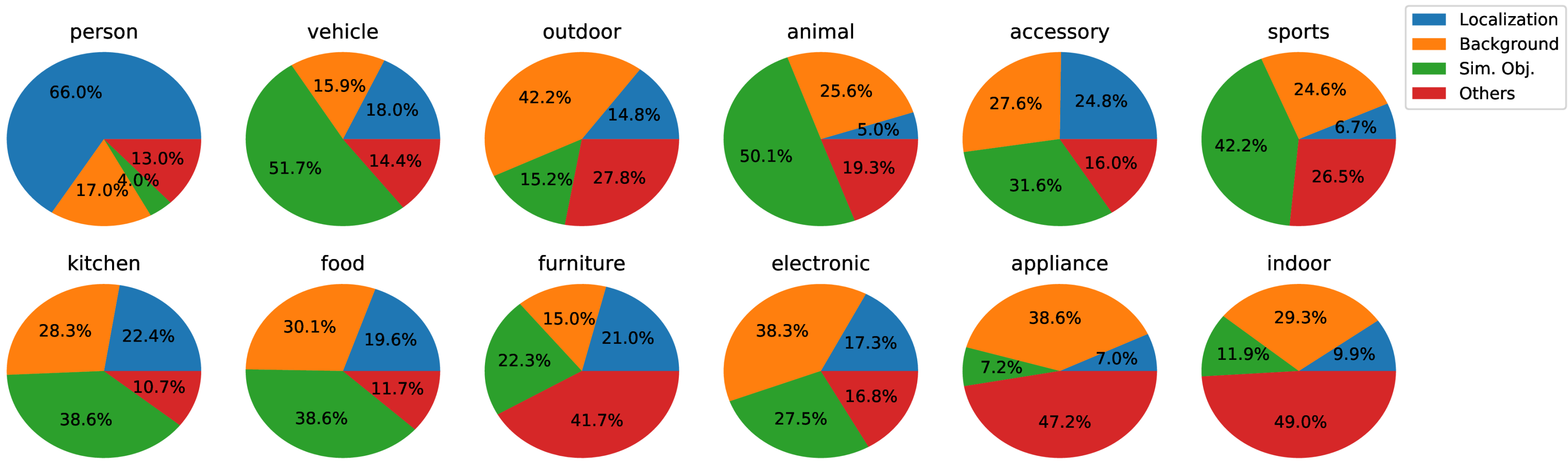


Figure 6.10: False positive analyses for superclasses on MS-COCO.

Table 6.6: mAP results on MS-COCO dataset with GZSD (65/15) settings.

Method	seen/unseen	seen	unseen	HM
Word embeddings	65/15	28.41	14.26	19.08
SimEmb	65/15	28.91	15.78	20.41

We present the results based on both the word embeddings directly and class-to-class similarities as class embeddings in Table 6.6. The results show that the standard word embedding scheme obtains 28.41% mAP on seen classes, 14.26% mAP on unseen classes and a harmonic mean score of 19.08. In contrast, the proposed embedding yields 28.91% 15.78% and 20.41 unseen mAP, seen mAP and harmonic mean scores, respectively. These results show that using class-to-class similarity vectors also provides a relative performance advantage in terms of model performance, while also enabling our effective alpha coefficient learning procedure.

6.4 Chapter Summary

An important shortcoming of current image captioning methods that aim training through non-paired datasets is that they do not work in a fully ZSL setting. These methods generate captions for images which consist of classes not seen in captioning datasets, but they assume that there is a ready-to-use fully supervised visual recognition model. To this end, we define the ZSIC problem, propose a novel GZSD model and a ZSIC approach based on it. We additionally introduce a practical class embedding scheme, a technique to improve GZSD performance via score scaling, and a novel evaluation method that provides insights into the ZSIC results. Our qualitative and quantitative experimental results show that our method yields promising results towards achieving our ZSIC goals. We believe that ZSIC is an important research direction towards building captioning models that are more suitable to use in realistic, in-the-wild settings.

CHAPTER 7

META-TUNING LOSS FUNCTIONS AND DATA AUGMENTATION FOR FEW-SHOT OBJECT DETECTION

7.1 Introduction

Object detection is one of the computer vision problems that has greatly benefited from the advances in supervised deep learning approaches. However, similar to the case in many other problems, state-of-the-art in object detection relies on the availability of large-scale fully-annotated datasets, which is particularly problematic due to the difficulty of collecting accurate bounding box annotations [**?**, 232]. This practical burden has lead to a great interest in the approaches that can potentially reduce the annotation cost, such as weakly-supervised learning [42, 41], learning from point annotations [233], and mixed supervised learning [43]. A more recently emerging paradigm in this direction is *few-shot object detection* (FSOD). In the FSOD problem, the goal is to build detection models for the *novel* classes with few labeled training

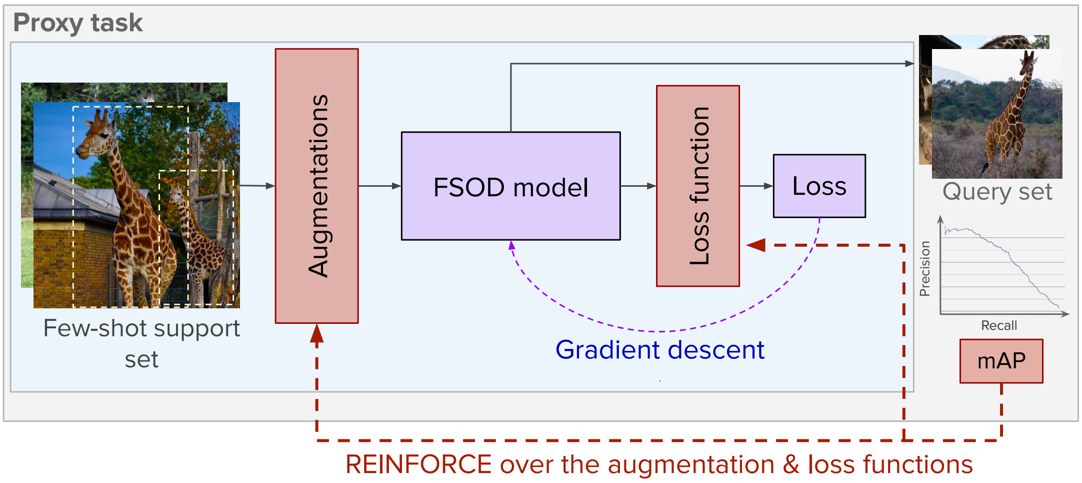


Figure 7.1: The overall architecture of the meta-tuning approach.

images by transferring knowledge from the *base* classes with a large set of training images. In the closely related Generalized-FSOD (G-FSOD) problem, the goal is to build few-shot detection models that perform well on both base and novel classes.

FSOD methods can be categorized into meta-learning and fine-tuning approaches. Although meta-learning based methods are predominantly used in the literature in FSOD research [60, 61, 62, 63, 64, 10, 65, 66, 67, 68], several fine-tuning based works have recently reported competitive results [11, 69, 70, 71, 72, 73, 74, 75]. The main premise of meta-learning approaches is to design and train dedicated meta-models that map given few train samples to novel class detection models, *e.g.* [234] or learn easy-to-adapt models [76] in a MAML [77] fashion. In contrast, however, fine-tuning based methods tackle the problem as a typical transfer learning problem and apply the general purpose supervised training techniques, *i.e.* regularized loss minimization via gradient-based optimization, to adapt a pre-trained model to few-shot classes. It is also worth noting that the recent results on fine-tuning based FSOD are aligned with related observations on few-shot classification [178, 187, 188] and segmentation [235].

While some of the FSOD meta-learning approaches are attractive for being able to learn dedicated parametric training mechanisms, they also come with two important shortcomings: (i) the risk of overfitting to the base classes used for training the meta-model due to model complexity, and (ii) the difficulty of interpreting what is actually learned; both of which can be crucially important for real-world, in-the-wild utilization of a meta-learned model. From this point of view, the simplicity and generality of a fine-tuning based FSOD approach can be seen as major advantages. In fact, one can find a large machine learning literature on the components (optimization techniques, loss functions, data augmentation, and architectures) of an FT approach, as opposed to the unique and typically unknown nature of a meta-learned inference model, especially when the model aims to replace standard training procedures for modeling the novel few-shot classes. While MAML [77] like meta-learning for quick adaptation is closer in nature to fine-tuning based approaches, the vanishing gradient problems and the overall complexity of the meta-learning task practically limits the approach to target only one or few model update steps, whereas an FT approach has no such computational difficulty.

Perhaps the biggest advantage of a fine-tuning based FSOD approach, however, can also be its biggest disadvantage: its generality may lack the inductive biases needed for effective learning with few novel class samples while preserving the knowledge of base classes. To this end, such approaches focus on the design of fine-tuning details, *e.g.* whether to freeze the representation parameters [11], use contrastive fine-tuning losses [69], increase the novel class variances [72], introduce the using additional detection heads and branches [70, 71]. However, optimizing such details specifically for few-shot classes in a hand-crafted manner is clearly difficult, and likely to be sub-optimal.

To address this problem, we focus on applying meta-learning principles to tune the loss functions and augmentations to be used in the fine-tuning stage for FSOD, which we call *meta-tuning* (Figure 7.1). More specifically, much like the meta-learning of a meta-model, we define an episodic training procedure that aims to progressively discover the optimal loss function and augmentation details for FSOD purposes in a data-driven manner. Using reinforcement learning (RL) techniques, we aim to tune the loss function and augmentation details such that they maximize the expected detection quality of an FSOD model obtained by fine-tuning to a set of novel classes. By defining meta-tuning over well-designed loss terms and an augmentation list, we restrict the search process to effective function families, reducing the computational costs compared to AutoML methods that aim to discover loss terms from scratch for fully-supervised learning [12, 197]. The resulting meta-tuned loss functions and augmentations, therefore, inject the learned FSOD-specific inductive biases into a fine-tuning based approach.

To explore the potential of the meta-tuning scheme for FSOD, we focus on the details of classification loss functions, based on the observations that FSOD prediction mistakes tend to be in classification rather than localization details [69]. In particular, we first focus on the softmax temperature parameter, for which we define two versions: (i) a simple constant temperature, and (ii) time (fine-tuning iteration index) varying dynamic temperature, parameterized as an exponentiated polynomial. In all cases, the parameters learned via meta-tuning yield an interpretable loss function that has a negligible risk of over-fitting to the base classes, in contrast to a complex meta-model. We also model augmentation magnitudes during meta-tuning for improving the data loading pipeline for few-shot learning purposes. Additionally, we incorporate a score scaling coefficient for learning to balance base versus novel class scores.

We provide an experimental analysis on the Pascal VOC [79] and MS-COCO [39] benchmarks for FSOD, using the state-of-the-art fine-tuning based baselines MPSR [70] and DeFRCN [75]. Our experimental results show that the proposed meta-tuning approach provides significant performance gains in both FSOD and Generalized FSOD settings, suggesting that meta-tuning loss functions and data augmentation can be a promising direction in FSOD research.

7.2 Method

This section provides a brief summary of the FSOD problem definition and the baseline model we utilize. We then present our definition and instantiation of meta-tuning.

Problem definition. We follow the FSOD setup of [10], where a relatively large set of training images for the set C_b of *base* classes is made available. Each training image corresponds to a tuple (x, y) consisting of image x and annotations $y = \{y_0, ..., y_M\}$. Each object annotation $y_i = \{c_i, b_i\}$ contains a category label (c_i) and a bounding box $(b_i = \{x_i, y_i, w_i, h_i\})$. Once the FSOD model training is complete, the evaluation is carried out based on a limited number (k) of training images made available for the set C_n of distinct *novel* (*i.e.* few-shot) classes.

Base model. We use the MPSR FSOD method [70] as the infrastructure for our loss function and data augmentation search methods. MPSR adapts the Faster-RCNN to be suitable for fine-tuning-based FSOD and uses an auxiliary multi-scale positive sample refinement (MPSR) branch to handle the scale scarcity problems. This branch expands the scale space of positive samples without increasing improper negative instances, unlike feature pyramid networks and image pyramids that do not change data distribution, hence the scale sparsity problem. In this context, objects in the images are cropped and resized in multiple sizes to create scale pyramids. The MPSR uses two groups of loss functions for the region proposal network (RPN) and detection heads, and feeds differently scaled positive samples to these loss functions together

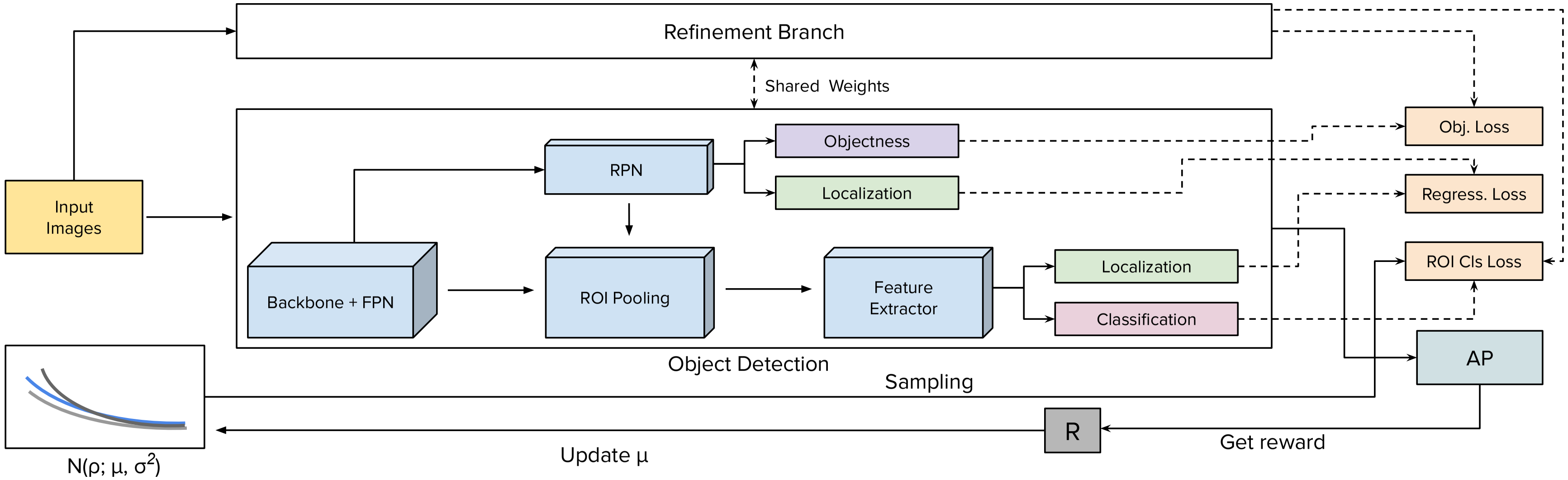


Figure 7.2: Details of our meta-tuning approach.

with the main detection branch. Finally, we note that the proposed approach can in principle be applied to virtually any fine-tuning based FSOD model.

7.2.1 Meta-tuning loss functions

Our main goal is to improve few-shot detector fine-tuning based on meta-learning principles. For meta-tuning the FSOD loss, we specifically focus on the classification loss term, as the FSOD errors tend to be primarily caused by misclassifications [69]. The MPSR classification loss term can be expressed as follows:

$$\ell_{cls}(x,y) = -\frac{1}{N_{ROI}} \sum_{i}^{N_{ROI}} \log\left(\frac{e^{f(x_i,y_i)}}{\sum_y e^{f(x_i,y)}}\right) \tag{7.1}$$

where N_{ROI} is the number of ROIs (*i.e.* candidate regions) in an image, y_i is the groundtruth class label for the i-th ROI, and $f(x_i, y)$ is the corresponding class y prediction score. To add more flexibility into the loss function, we re-define it as a parametric function $\ell_{cls}(x, y; \rho)$, where ρ represents the loss function parameters. First, we introduce a temperature scalar ρ_τ, *i.e.* $\rho = (\rho_\tau)$:

$$\ell_{cls}(x,y;\rho) = -\frac{1}{N_{ROI}} \sum_{i}^{N_{ROI}} \log\left(\frac{e^{f(x_i,y_i)/\rho_\tau}}{\sum_{y'} e^{f(x_i,y')/\rho_\tau}}\right) \tag{7.2}$$

Our motivation comes from the observations on the importance of temperature scaling in log loss on various other problems, such as knowledge distillation [236], few-shot classification [172, 171], and zero-shot learning [4]. While temperature is typically tuned in a manual manner, here we aim to meta-learn it specifically for fine-tuning based FSOD purposes, giving a chance to observe the behavior of meta-tuning in a simple case. We also define a more sophisticated variant of the loss function by defining the *dynamic temperature* function f_ρ and *novel class scaling* α:

$$\ell_{cls}(x,y;\rho) = \frac{-1}{N_{ROI}} \sum_{i}^{N_{ROI}} \log\left(\frac{e^{\alpha(y_i)f(x_i,y_i)/f_\rho(t)}}{\sum_{y'} e^{\alpha(y')f(x_i,y')/f_\rho(t)}}\right) \tag{7.3}$$

where $f_\rho(n) = \exp(\rho_a n^2 + \rho_b n + \rho_c)$. Here, $\rho = (\rho_a, \rho_b, \rho_c)$ is a 3-tuple of polynomial coefficients, and $n \in [0, 1]$ is the normalized fine-tuning iteration index. The temperature can increase or decrease over time, making the predicted class distributions smoother or sharper. $\alpha(y)$ is set to 1 for $y \in C_b$, and otherwise the novel class score scaling coefficient ρ_α, as a way to learn base and novel score balancing.

7.2.2 Meta-tuning augmentations

For meta-tuning augmentations, we focus on the photometric augmentations that are likely to be transferable from base to novel classes. In this context, we model the brightness, saturation, contrast, and hue transforms, with a shared magnitude parameter (ρ_{aug}), which is known to be effective for supervised training [202].

7.2.3 Meta-tuning procedure

In our work, we utilize a REINFORCE [237] based reinforcement learning (RL) approach to search for the optimal loss function and augmentations, where we use the AutoML approach of Wang *et al.* [199] on loss function search for fully-supervised face recognition as our starting point.

In order to meta-tune the loss function and augmentations to maximize FSOD generalization abilities, we generate *proxy tasks* over base class training data to imitate real FSOD tasks over the novel classes. For this purpose, we divide base classes into two subsets, proxy-base $C_{\text{p-base}}$ and proxy-novel $C_{\text{p-novel}}$. We then construct three non-overlapping data set splits using the base class training set: (i) $D_{\text{p-pretrain}}$ containing $C_{\text{p-base}}$-only samples, used for training a temporary object detection model for meta-tuning purposes; (ii) $D_{\text{p-support}}$ containing samples of $C_{\text{p-base}} \cup C_{\text{p-novel}}$ classes to be used as fine-tuning images during meta-tuning; (iii) $D_{\text{p-query}}$ containing samples of $C_{\text{p-base}} \cup C_{\text{p-novel}}$ classes to be used for evaluating the generalized FSOD performance of a fine-tuned model during meta-tuning.

We generate a series of FSOD proxy tasks for meta-tuning, similar to episodic meta-learning: at each proxy task T, we sample a few-shot training set from $D_{\text{p-support}}$. We also sample a loss function/augmentation magnitude parameter combination ρ, where each $\rho_j \in \rho$ is modeled in terms of a Gaussian distribution: $\rho_j \sim \mathcal{N}(\mu_j, \sigma^2)$. Using the loss function or augmentations corresponding to the sampled ρ, we fine-tune the initial model on the support images using gradient-based optimization, and compute the mean average precision (mAP) scores on $D_{\text{p-query}}$. We get multiple mAP scores by repeating this process multiple times over multiple proxy support samples. Meta-tuning is then carried over by updating μ values via the REINFORCE rule after each episode, towards finding μ values centered around well-performing ρ combinations.

$$\mu_j' \leftarrow \mu_j + \eta R(\rho) \nabla_\mu \log\left(p(\rho_j; \mu_j, \sigma)\right) \tag{7.4}$$

where $p(\rho; \mu, \sigma)$ is the Gaussian probability density function, η is the RL learning rate.

We apply the REINFORCE update rule using the ρ with the highest reward per episode. $R(\rho)$ is the *normalized* reward function obtained by whitening the mAP scores. We empirically observe that normalization improves the results (Section 7.3) since without reward normalization, the RL updates are scaled with respect to the inherent difficulty of the proxy task, which greatly varies depending on the sampled support examples. Reward normalization approximately removes the *average* reward, enabling better performing ρ samples to influence based on their *relative* success.

Finally, similar to [238], starting with $\sigma = 0.1$, we diminish σ over the RL iterations to progressively reduce explorations by sampling more conservatively, which improves converge. The final scheme is illustrated in Figure 7.2.

7.3 Experiments

Metrics. We use mAP to evaluate the base and novel class detection results separately. To evaluate the generalized FSOD performance, we use the Harmonic Mean (HM) metric to compute a balanced aggregation of base and novel class performance scores. Adapted from generalized zero-shot learning [239], HM is defined as the harmonic mean of mAP_{base} and $\text{mAP}_{\text{novel}}$ scores.

Datasets. We use Pascal VOC [79] and MS COCO [39] with the same splits defined in FSOD benchmarks [11, 70]. On Pascal VOC, three separate base/novel class splits exist, where each one consists of 15 base and 5 novel classes. In each split, we select 5 base classes to mimic novel classes during meta-tuning. On MS-COCO, we select 15 base classes to mimic novel classes in each proxy task, and evaluate the models for the 10-shot and 30-shot settings.

Baselines. We primarily use the MPSR [70] and DeFRCN [75] as our baselines, which are among the best performing fine-tuning based FSOD methods on Pascal VOC. For

Table 7.1: FSOD and G-FSOD results on Pascal VOC and MS-COCO datasets for MPSR baseline method.

Method/Shot	Pascal VOC										MS-COCO			
	Novel Classes					All Classes (HM)					Novel Classes		All Classes (HM)	
	1	2	3	5	10	1	2	3	5	10	10	30	10	30
MPSR [70]	33.1	37.2	44.3	47.1	52.1	43.1	47.4	54.5	57.2	60.8	9.1	13.7	11.5	15.0
MPSR+Meta-Static	33.4	39.4	45.1	47.3	52.6	43.7	50.4	55.4	57.5	61.4	10.1	14.8	12.7	16.4
MPSR+Meta-Dynamic	34.5	39.8	45.0	48.2	52.5	45.0	51.0	55.5	58.3	61.6	11.9	14.9	14.3	16.6
MPSR+Meta-ScaledDynamic	35.2	40.3	45.8	48.4	52.9	45.6	51.2	55.9	58.3	61.8	12.3	15.0	14.4	16.7
MPSR+Aug	34.6	38.6	46.0	48.3	52.7	45.1	49.5	56.2	58.4	62.0	9.9	14.9	12.5	16.3
MPSR+Meta-Static+Aug	35.3	39.1	46.1	48.4	52.7	45.9	49.9	56.2	58.3	61.8	10.2	15.2	12.8	16.7
MPSR+Meta-Dynamic+Aug	35.4	39.6	46.5	48.9	53.3	46.0	50.5	56.8	58.9	62.5	12.1	15.3	14.5	16.8
MPSR+Meta-ScaledDynamic+Aug	**35.8**	**40.6**	**46.8**	**49.2**	**53.7**	**46.3**	**51.5**	**57.0**	**59.2**	**62.7**	**12.5**	**15.4**	**14.7**	**16.9**

the DeFRCN experiments, we transfer the meta-tuned loss functions and augmentation magnitudes from MPSR to the DeFRCN method, which are both based on Faster-RCNN. We take the results for FRCN [64], Ret. R-CNN [71], Meta-RCNN [64], FSRW [10], MetaDet [240], FsDetView [60] and ONCE [65] from [71] for a fair comparison. For the MPSR, DeFRCN (*seed* is set to 0) and FSCE [69], we report the results we obtain experimentally. We take the results for TFA+Hal [72], CME [62], TIP [194], DCNet [192], QA-FewDet [193] FADI [73], LVC [74], KFSOD [67] and FCT [68] from the original papers. Finally, while it is difficult to fairly compare fine-tuning versus meta-learning based approaches, we provide a discussion in the supplementary material.

Implementation details. We use 200 RL episodes for loss function meta-tuning, with REINFORCE learning rate set to 0.0005. The meta-tuning for augmentation parameter is carried out using the trained and frozen the loss function parameters. We keep the fine-tuning implementation details of MPSR unchanged, which uses 4000 and 8000 gradient descent iterations for 10-shot and 30-shot experiments on MS-COCO, and 2000 iterations on Pascal VOC. We will publish the full source code upon publication; a preliminary version is provided as supplementary material.

7.3.1 Main results

We first compare the meta-tuning results against the corresponding MPSR baseline in Table 7.1. In the table, *Meta-Static*, *Meta-Dynamic*, *Meta-ScaledDynamic* refer to meta-tuning a single temperature, dynamic temperature, and novel class scaled dynamic temperature functions, respectively. Similarly, *Aug*, *Meta-Static+Aug*, *Meta-Dynamic+Aug*, and *Meta-ScaledDynamic+Aug* refer to meta-tuning only augmentation, single temperature and augmentation, dynamic temperature and augmentation, and novel class scaled dynamic temperature and augmentation functions, respectively. We observe that meta-tuning consistently improves the FSOD and G-FSOD results of the MPSR model. We also observe steady improvements gradually from the baseline to Meta-Static, to Meta-Dynamic, and finally to Meta- ScaledDynamic. In addition, the meta-tuned augmentation magnitude parameter also contributes positively to the few-shot object detection performance. The overall consistency of the improvements provides positive evidence for the value of loss and augmentation meta-tuning.

Pascal VOC results. In Table 7.2, we report the Pascal VOC results for our MPSR and DeFRCN based Meta-ScaledDynamic+Aug approach and compare them against the state-of-the-art fine-tuning based FSOD methods. In this table, the best and the second-best results are marked with red and blue colors, respectively. While we present the scores averaged over the three splits in this table, additional per-split FSOD and G-FSOD results can be found in the supplementary material. The left side of Table 7.2 presents the FSOD results for the varying number of support images. We observe that DeFRCN combined with Meta-ScaledDynamic+Aug, *i.e.* meta-tuning of the score coefficient, dynamic temperature and the augmentation parameter, yields the best mAP scores in all k-shot settings among all methods.

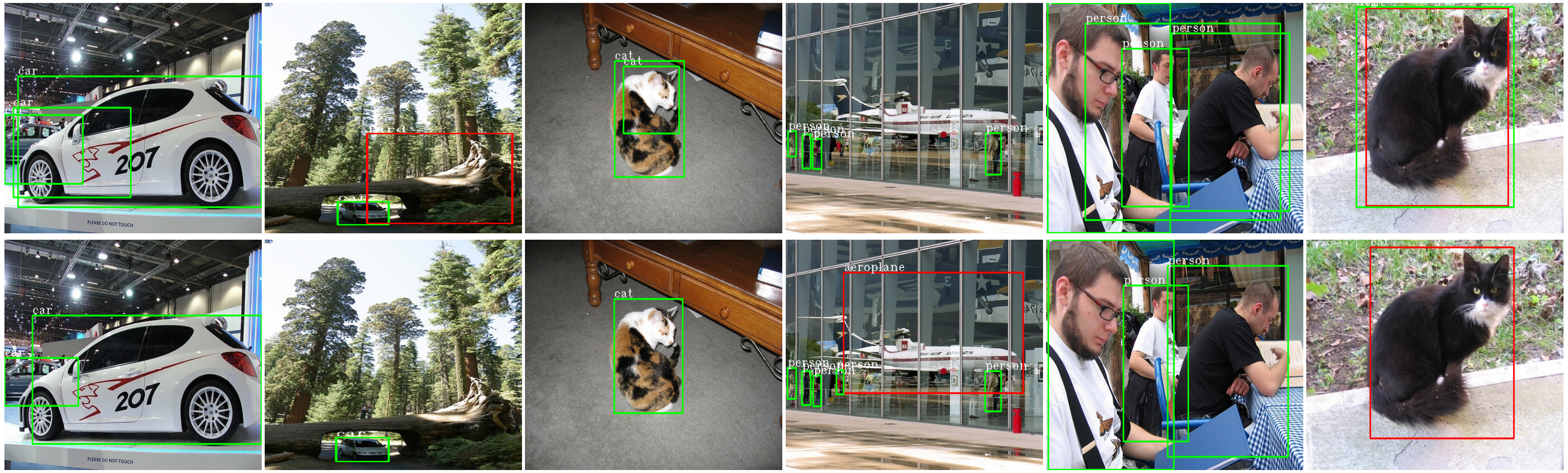
The right side of Table 7.2 presents the G-FSOD results on Pascal VOC. We observe that the best-performing Meta-ScaledDynamic+Aug method improves the HM scores further above the state-of-the-art in all k-shot settings. Overall, these results suggest that the proposed framework is an effective way for meta-learning inductive biases to be used in fine-tuning-based FSOD.

Figure 7.3 presents visual detection examples without and with meta-tuned scaled dynamic temperature and augmentations in the first and second rows, respectively. In this figure, base and novel class detections are shown with green and red boxes. We observe various improvements, such as reductions in false positives, improved recall, and more precise boxes, most likely due to the improved model fitting in the low-data regime.

MS-COCO results. In Table 7.3, we compare the MPSR and DeFRCN based Meta-ScaleDynamic+Aug results against other fine-tuning based FSOD methods that report 10-shot and 30-shot results on the MS-COCO dataset. We observe that with meta-tuning, the FSOD scores of MPSR improve from 9.1 to 12.5 (10-shot mAP), and from 13.7 to 15.4 (30-shot mAP). We also observe that the scores of DeFRCN improve from 18.5 to 18.8 (10-shot mAP), and from 21.9 to 23.4 (30-shot mAP), obtaining the best and second best results against all other models. Similarly, in the case of G-FSOD, with meta-tuning, the 10-shot HM score of DeFRCN improves from 24.0 to

Table 7.2: FSOD (mAP) and G-FSOD (HM of the base and novel class mAPs) results on Pascal VOC.

Method/Shot	Novel Classes					All Classes (HM)				
	1	2	3	5	10	1	2	3	5	10
FRCN [64] (ICCV'19)	16.1	20.6	28.8	33.4	36.5	25.9	31.7	40.0	44.3	46.7
TFA-fc [11] (ICML'20)	27.6	30.6	39.8	46.6	48.7	40.5	44.1	52.9	58.3	59.9
TFA-cos [11] (ICML'20)	31.4	32.6	40.5	46.8	48.3	44.6	46.0	53.5	58.4	59.6
FSCE [69] (CVPR'21)	29.2	36.3	42.5	47.1	52.2	41.8	48.8	54.2	57.7	61.0
Ret. R-CNN [71] (CVPR'21)	31.4	37.1	41.4	46.8	48.8	44.7	50.5	54.7	59.1	60.8
TFA+Hal [72] (CVPR'21)	32.9	35.5	40.4	46.3	48.1	-	-	-	-	-
FADI [73] (NeurIPS'21)	42.2	46.5	47.9	52.4	56.9	-	-	-	-	-
LVC [74] (CVPR'22)	30.9	35.4	43.6	51.1	54.1	-	-	-	-	-
LVC-PL [74] (CVPR'22)	45.2	45.0	54.8	57.5	58.6	-	-	-	-	-
MPSR [70] (ECCV'20)	33.1	37.2	44.3	47.1	52.1	43.1	47.4	54.5	57.2	60.8
DeFRCN [75] (ICCV'21)	**46.5**	**52.6**	**55.9**	**60.0**	**60.8**	**57.6**	**62.5**	**64.7**	**67.6**	**67.8**
MPSR+Meta-ScaledDynamic+Aug	35.8	40.6	46.8	49.2	53.7	46.3	51.5	57.0	59.2	62.7
DeFRCN+Meta-ScaledDynamic+Aug	**49.2**	**54.0**	**57.2**	**61.3**	**61.8**	**59.8**	**63.7**	**65.9**	**68.6**	**68.7**



Split-1/5-shot Split-1/10-shot Split-2/5-shot Split-2/10-shot Split-3/5-shot Split-3/10-shot

Figure 7.3: Qualitative Pascal VOC results using MPSR without (first row) and with (second row) meta-tuning.

Table 7.3: Comparison of Meta-ScaledDynamic results to the fine-tuning based GFSOD methods on the MS-COCO dataset.

Method/Shots	Novel Classes		All Classes (HM)	
	10-shot	30-shot	10-shot	30-shot
FRCN [64] (ICCV'19)	9.2	12.5	12.8	15.6
FRCN-BCE [11] (ICML'20)	6.4	10.3	10.9	16.1
TFA-fc [11] (ICML'20)	10.0	13.4	15.4	19.4
TFA-cos [11] (ICML'20)	10.0	13.7	15.6	19.8
MPSR [70] (ECCV'20)	9.1	13.7	11.5	15.0
FSCE [69] (CVPR'21)	10.5	14.4	16.0	20.2
Ret. R-CNN [71] (CVPR'21)	10.5	13.8	16.6	20.4
FADI [73] (NeurIPS'21)	12.2	16.1	-	-
DeFRCN [75] (ICCV'21)	**18.5**	21.9	**24.0**	26.8
LVC [74] (CVPR'22)	12.1	17.8	17.8	22.8
LVC-PL [74] (CVPR'22)	17.8	**24.5**	22.8	**28.1**
MPSR+Meta-ScaledDynamic+Aug	12.5	15.4	14.7	16.9
DeFRCN+Meta-ScaledDynamic+Aug	**18.8**	**23.4**	**24.4**	**28.0**

24.4, outperforming all other models. In addition, the 30-shot HM score of DeFRCN improves from 26.8 to 28.0, which is slightly below the 28.1 score of LVC-PL [74].

7.3.2 Ablation studies

Meta-tuning details. The proposed meta-tuning approach involves three important technical details: *Proxy-novel imitation*, *model re-initialization*, and *reward normalization*. Proxy-novel imitation refers to reinforcement learning over the sampled proxy-novel tasks, instead of the whole training set, to mimic the test-time FSOD challenges. Model re-initialization is the re-initialization of the base model for each task. Without re-initialization, not only the sampled loss/augmentation parameters and tasks but also the accumulated model updates undesirably affect the rewards. Reward

Table 7.4: Evaluation of meta-tuning details.

Proxy-novel imit.	Model re-init.	Reward norm.	HM
✗	✗	✗	61.5
✓	✗	✗	61.8
✓	✓	✗	62.1
✓	✓	✓	63.3

normalization further reduces the effect of task difficulty variance by normalizing the rewards obtained within a single episode, allowing a more isolated assessment of the sampled loss functions and augmentations.

We evaluate the contributions of these three important details in terms of G-FSOD HM scores using the 5-shot setting of Pascal VOC Split-1 with MPSR+Meta-Dynamic. The results averaged over 5 runs are given in Table 7.4, where *proxy-novel imitation* is the imitation of novel classes using a subset of base classes, *model re-initialization* is the re-initialization of the base model at each task, and *reward normalization* is within-episode normalization of the mAP scores during meta-tuning. We observe that each component progressively improves the HM scores, and the most significant contribution is made by reward normalization, which improves from 62.1 to 63.3. We also observe that reward normalization considerably improves the overall experimental stability. To quantify this observation, we estimate the 95% confidence interval over the runs using $CI = 1.96\frac{s}{\sqrt{n}}$, where s, n, and 1.96 are the standard deviation, number of runs, and Z-value, respectively [11]. According to this estimator, the normalization step narrows the confidence interval from ± 0.75 to ± 0.13, providing a clear improvement in reliability.

Learned loss functions. In Figure 7.4, we plot the learned loss functions according to the μ values obtained at the end of the RL process. The upper plot shows the dynamic temperature functions learned over three different splits. We observe that temporally attenuated temperature values are preferred consistently, sharpening the predictions towards the end of the fine-tuning process. The lower plot shows the learned dynamic temperature functions with novel class score scaling. The learned scaling coefficients,

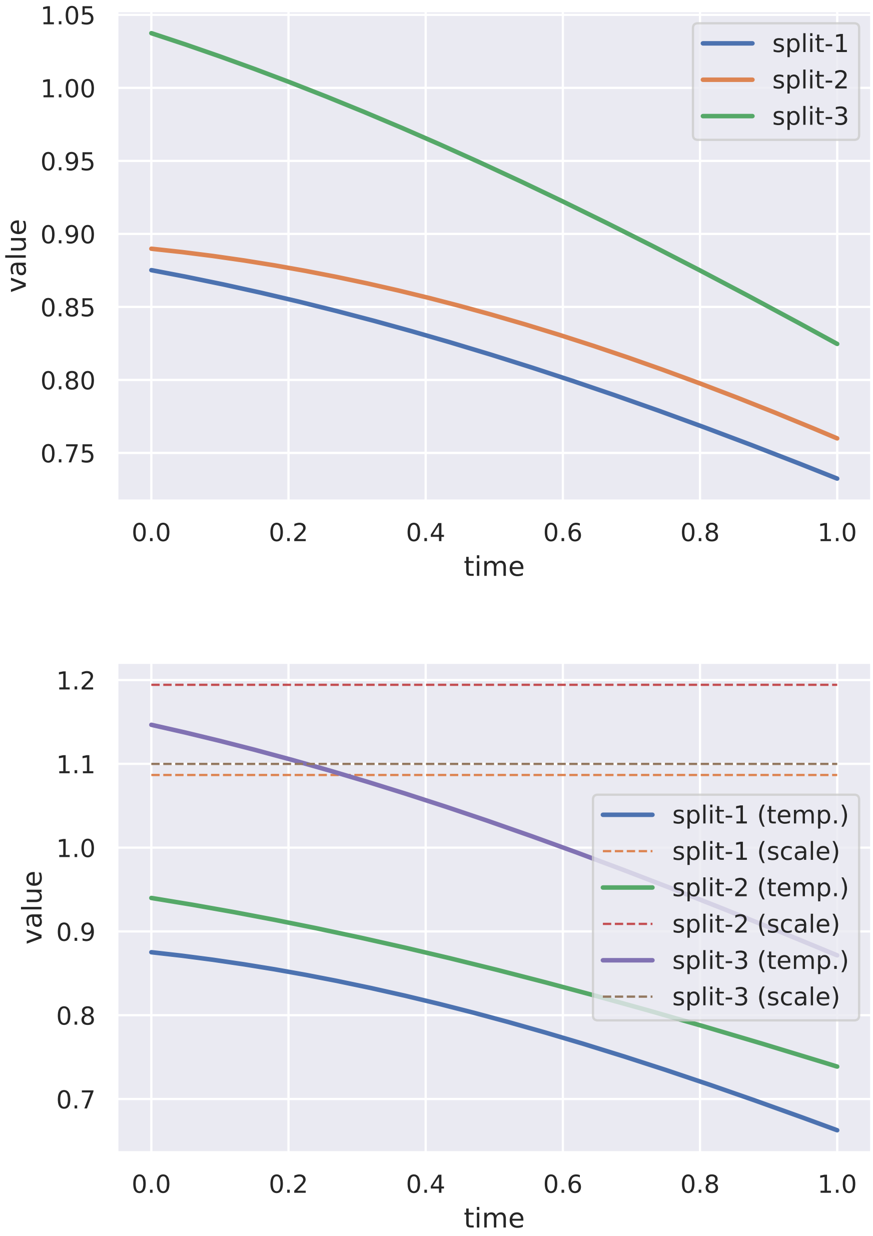


Figure 7.4: The dynamic temperature functions and score scaling coefficients learned by the meta-tuning process.

i.e. μ_α of the learned ρ_α distribution, are shown as horizontal lines. We observe that similar dynamic temperature functions are learned, and μ_α values vary between 1.09 to 1.2, suggesting that the meta-tuning process learns to boost the novel class scores. The interpretability of these outcomes, we believe, highlights a significant advantage of loss meta-tuning. In the context of interpretability, we observe that as

Table 7.5: Low-shot (1-shot, 2-shot and 3-shot) experiments on MS-COCO dataset with novel classes.

S/M	TFA [11]	TFA+Hal [72]	TFA+Meta-ScaledDynamic+Aug
1	3.4	3.8	**4.7**
2	4.6	5.0	**5.8**
3	6.6	6.9	**7.1**

the fine-tuning process continues on the few-shot training set, the predictions are progressively made sharper, *i.e.* the loss becomes more sensitive to classification errors and enforces towards making more confident correct predictions. This is in alignment with one of our original motivations for reducing the dominating classification errors in G-FSOD, as the meta-tuning process automatically learns to enforce more accurate classifications, where the curve steepness and the numerical ranges are learned via RL.

Learned augmentations. The learned photometric augmentation magnitude values learned are 0.29, 0.24, 0.13, and 0.36 for Pascal VOC split-1, split-2, split-3, and MS-COCO datasets, respectively. We observe that the learned augmentation magnitudes positively contribute to the performance. According to the results in Table 7.1, the average Pascal VOC split-1/1-shot score increases from 33.1 to 34.6 with only augmentation steps.

Very low-shot experiments. Finally, we evaluate the meta-tuning approach in low-shot many-class settings. [72] proposes TFA+Hal method that uses the TFA baseline and conducts 1-shot, 2-shot, and 3-shot FSOD on the MS-COCO dataset. As we already observe the positive effects of the loss terms and augmentation magnitudes obtained from the MPSR on the DeFRCN, we similarly apply the learned parameters to the TFA baseline. The results are presented in Table 7.5. We observe that results are consistently improved using the meta-tuned functions on the TFA baseline.

7.4 Chapter Summary

Fine-tuning based frameworks offer simple and reliable approaches to building detection models from few samples. However, a major limitation of the existing fine-tuning-based FSOD models is their focus on the hand-crafting the design of fine-tuning details for few-shot training, which is inherently difficult and likely to be sub-optimal. Towards addressing this limitation, we propose to meta-learn the fine-tuning based learning dynamics as a way of introducing learned inductive biases for few-shot learning. The proposed tuning scheme uses meta-learning principles with reinforcement learning, and obtains interpretable loss functions and augmentation magnitudes for few-shot training. Our comprehensive experimental results on Pascal VOC and MS COCO datasets show that the proposed meta-tuning approach consistently provides significant performance improvements over the strong fine-tuning based few-shot detection baselines in both FSOD and G-FSOD settings.

7.5 More on Meta-tuning

In this section, we give some details of our work. In this context, we share the class splits used in proxy tasks in Section 7.5.1, the meta-tuning algorithm in Section 7.5.2, additional experimental results which belong to various fine-tuning and meta-tuning based FSOD and G-FSOD methods in Section 7.5.3, and implementation runtime information in Section 7.5.4, respectively. Moreover, we also share some randomly sampled visual results of *MPSR+Meta-ScaledDynamic*, *MPSR+Meta-ScaledDynamic+Aug* and *DeFRCN+Meta-ScaledDynamic+Aug* methods.

7.5.1 Proxy task class splits

We use proxy tasks to apply the meta-tuning ideas, so we generate sub-splits in the base classes. In this context, we select some base classes to mimic novel classes to conduct the proxy task. We summarize the list of proxy Pascal VOC classes on Table 7.6. The list of selected proxy novel classes for the MS-COCO dataset is as follows:

{*"skis", "tennis racket", "scissors", "truck", "baseball bat", "handbag", "carrot", "mouse", "parking meter", "apple", "knife", "microwave", "refrigerator", "cake", "zebra"*}.

Table 7.6: Proxy task class splits for Pascal VOC.

Proxy-base classes ($C_{\textbf{p-base}}$)			**Proxy-novel classes** ($C_{\textbf{p-novel}}$)		
Split-1	Split-2	Split-3	Split-1	Split-2	Split-3
aeroplane	bicycle	aeroplane	person	motorbike	horse
bicycle	bird	bicycle	pottedplant	person	person
boat	boat	bird	sheep	sheep	pottedplant
bottle	bus	bottle	train	train	train
car	car	bus	tvmonitor	tvmonitor	tvmonitor
cat	cat	car			
chair	chair	chair			
diningtable	diningtable	cow			
dog	dog	diningtable			
horse	pottedplant	dog			

7.5.2 Algorithm

We summarize the main meta-tuning procedure in Algorithm 1. We can divide this algorithm into three parts: (i) model initialization and parameter sampling, (ii) instance sampling and mAP calculation, (iii) mAP normalization and RL steps.

1) Model initialization and parameter sampling. This algorithm firstly initializes the base proxy detection model weights for the proxy task and sample ρ value from normal distributions. The base proxy detection model represents the object detection model trained using the $D_{\text{p-pretrain}}$ dataset.

2) Instance sampling and mAP calculation. The proposed algorithm samples new instances from the proxy fine-tuning dataset $D_{\text{p-support}}$, and calculates the mean average precision scores on proxy validation dataset $D_{\text{p-query}}$ after a certain number of iterations. The algorithm repeats this process for N times.

3) mAP normalization and RL steps. The proposed algorithm normalizes the mAP scores, selects the maximum score as the reward value among the normalized APs, and applies a single REINFORCE step.

Algorithm 1 Meta-tuning Loss Function Learning

```
Input: Pre-trained model $m_{init}$, proxy fine-tuning dataset $D_{\text{p-support}}$, proxy validation dataset $D_{\text{p-query}}$, number of $rho$ trials $N$, maximum iteration number $M$

iteration_index = 1
repeat
    Initialize $m_{init}$ and sample new $\rho$

    for $rho\_index = 1$ to $N$ do
        Sample new fine-tuning images from $D_{\text{p-support}}$
        Take $m_{init}$, run all iter. using current samples
        Calculate mAP[$rho\_index$] on $D_{\text{p-query}}$
    end for

    Normalize mAP scores
    Get max normalized AP as a reward
    Make a single REINFORCE step
    iteration_index += 1
until iteration_index = M
```

7.5.3 Additional Experimental Results

In this section, we share detailed experimental comparison results for Pascal VOC and MS COCO datasets.

Comparison to fine-tuning based FSOD and G-FSOD methods on Pascal VOC. We first present the detailed Pascal VOC comparisons for each split and shot with only novel classes in Table 7.7, and the detailed comparisons with all classes in Table 7.8. The experimental results show that the meta-tuning approach significantly improves the strong fine-tuning based few-shot detection baselines on the Pascal VOC benchmark. We provide complementary visual results of MPSR+Meta-ScaledDynamic and MPSR+Meta-ScaledDynamic+Aug methods using the Pascal VOC split-3/10-shot setting in Figure 7.5 and Figure 7.6, respectively. We also present examples from the visual results of the DeFRCN+Meta-ScaledDynamic+Aug method using the Pascal VOC split-2/10-shot setting in Figure 7.7. In these figures, base class instance candidates are marked with green, and novel class instance candidates are marked with red color.

Comparisons to meta-learning based FSOD and G-FSOD on Pascal VOC. We present the detailed Pascal VOC comparisons with meta-learning based methods in Table 7.9 and Table 7.10 for novel-only and all-classes settings, respectively. Since the most of the meta-learning methods do not share G-FSOD results, we are able to compare against a more limited number of meta-learning methods than FSOD. The experimental results (Table 7.9) show that our DeFRCN+Meta-ScaledDynamic+Aug method obtains the best results in all of the FSOD cases, except for the Split-2/1-shot setting. In the G-FSOD experiments (Table 7.10), it is observed that the proposed meta-tuning approach obtains the state-of-the-art results with a clear margin against existing meta-learning based methods.

Comparisons to meta-learning based FSOD and G-FSOD on MS-COCO. We compare our results with meta-learning based methods on the MS-COCO dataset and share the obtained results in Table 7.11. In this table, we are able to report a rather limited number of meta-learning methods to compare the G-FSOD results since most meta-learning based methods do not share G-FSOD results on the MS-COCO dataset. In FSOD experiments, we also observe that our DeFRCN+Meta-ScaledDynamic+Aug method obtains higher results than several recently published meta-learning based methods. We additionally observe major improvements in terms of HM scores in the G-FSOD setting, similar to the improvements obtained on the Pascal VOC dataset.

7.5.4 Implementation and runtime

We run our MPSR and DeFRCN experiments on a server with 4 Nvidia Tesla V100 32GB GPUs. The base MPSR model training to be used during fine-tuning takes 0.25 days for Pascal VOC and 0.45 days for MS COCO datasets. Since the base models used for the proxy tasks contain fewer classes and demand fewer iterations, the training of the MPSR model takes 0.1 days in Pascal VOC and 0.6 days in MS COCO datasets for the proxy-base classes. RL training for meta-tuning using the final setting takes 0.05 days for Pascal VOC splits and 0.5 days for the MS COCO dataset. Finally, we note that meta-tuning operations do not incur any overhead during the fine-tuning for novel classes.

Table 7.7: Comparison to fine-tuning based FSOD methods on the Pascal VOC dataset, with only novel classes.

Method/Shot	Split 1					Split 2					Split 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FRCN [64] (ICCV'19)	15.2	20.3	29.0	25.5	28.7	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
TFA-fc [11] (ICML'20)	36.8	29.1	43.6	55.7	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2
TFA-cos [11] (ICML'20)	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR [70] (ECCV'20)	37.2	43.6	50.9	53.7	60.2	24.8	28.1	38.0	39.8	45.9	37.3	40.0	43.9	47.8	50.1
Ret. R-CNN [71] (CVPR'21)	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1
TFA+H [72] (CVPR'21)	45.1	44.0	44.7	55.0	55.9	23.2	27.5	35.1	34.9	39.0	30.5	35.1	41.4	49.0	49.3
FSCE [69] (CVPR'21)	37.6	44.7	46.9	52.2	60.3	24.5	30.1	38.2	40.4	45.9	25.4	34.2	42.3	48.7	50.3
FADI [73] (NeurIPS'21)	50.3	54.8	54.2	59.3	63.2	30.6	35.0	40.3	42.8	48.0	45.7	49.7	49.1	55.0	59.6
LVC [74] (CVPR'22)	36.0	40.1	48.6	57.0	59.9	22.3	22.8	39.2	44.2	47.8	34.3	43.4	42.9	52.0	54.5
LVC-PL [74] (CVPR'22)	**54.5**	53.2	58.8	63.2	65.7	**32.8**	29.2	**50.7**	49.8	50.6	48.4	52.7	55.0	59.6	59.6
DeFRCN [75] (CVPR'21)	53.7	**59.5**	**61.2**	**65.7**	**66.6**	32.3	**42.0**	49.5	**52.4**	**53.4**	**53.6**	**56.2**	**56.9**	**61.9**	**62.3**
MPSR+Meta-Static	36.7	47.0	52.1	53.8	60.8	25.3	31.6	38.4	40.8	46.9	38.3	39.7	44.8	47.2	50.1
MPSR+Meta-Dynamic	40.4	47.5	51.9	54.9	60.5	25.6	31.7	38.5	40.6	46.7	37.6	40.2	44.7	49.1	50.3
MPSR+Meta-ScaledDynamic	41.5	47.9	52.7	55.4	60.9	25.7	32.2	38.9	40.8	46.8	38.5	40.9	45.9	49.0	51.0
MPSR+Aug	39.5	47.1	53.2	54.9	59.5	26.2	31.0	39.7	41.8	47.8	38.0	37.8	45.2	48.4	50.9
MPSR+Meta-Static+Aug	40.9	47.6	53.6	54.7	60.2	26.5	31.6	38.9	42.2	47.3	38.7	38.1	45.8	48.2	50.8
MPSR+Meta-Dynamic+Aug	41.0	47.5	53.8	55.2	60.2	26.4	32.2	39.8	42.7	48.5	38.9	39.1	46.0	48.8	51.3
MPSR+Meta-ScaledDynamic+Aug	41.8	48.7	54.2	55.7	61.1	26.5	32.7	40.0	42.5	48.7	39.0	40.4	46.2	49.6	51.2
DeFRCN+Meta-ScaledDynamic+Aug	**58.4**	**62.4**	**63.2**	**67.6**	**67.7**	**34.0**	**43.1**	**51.0**	**53.6**	**54.0**	**55.1**	**56.6**	**57.3**	**62.6**	**63.7**

Table 7.8: Comparison to fine-tuning based G-FSOD methods on the Pascal VOC dataset, with both base and novel classes.

Method/Shot	Split-1					Split-2					Split-3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FRCN [64] (ICCV'19)	24.9	31.4	40.3	37.6	41.0	22.1	31.3	39.1	43.0	47.5	30.8	32.3	40.5	52.2	51.7
TFA-fc [11] (ICML'20)	50.4	42.6	56.2	65.4	66.1	29.7	42.4	47.0	49.0	52.1	41.3	47.4	55.6	60.6	61.6
TFA-cos [11] (ICML'20)	53.1	49.5	57.1	65.4	65.3	36.3	40.0	47.6	48.6	52.2	44.5	48.5	55.9	61.2	61.4
MPSR [70] (ECCV'20)	45.8	52.5	59.3	61.8	65.5	36.0	39.7	49.8	51.7	56.9	47.6	49.9	54.5	58.1	60.0
FSCE [69] (CVPR'21)	50.7	56.5	58.1	61.6	66.1	36.5	42.4	49.8	51.5	55.8	38.2	47.4	54.6	59.9	61.1
Ret. R-CNN [71] (CVPR'21)	55.6	58.5	58.6	64.5	66.2	34.3	41.5	49.2	51.0	54.0	44.1	51.6	56.4	61.9	62.2
DeFRCN [75] (CVPR'21)	**63.3**	**67.3**	**68.1**	**71.1**	**71.2**	**45.9**	**54.7**	**60.3**	**62.8**	**63.1**	**63.7**	**65.4**	**65.5**	**68.8**	**69.2**
MPSR+Meta-Static	45.7	56.4	60.3	62.1	66.1	36.7	43.7	50.3	52.7	57.9	48.6	51.2	55.5	57.8	60.1
MPSR+Meta-Dynamic	50.2	57.2	60.6	63.3	67.0	37.0	43.9	50.4	52.5	57.8	47.9	51.8	55.4	59.1	60.2
MPSR+Meta-ScaledDynamic	51.0	57.3	60.9	63.3	67.1	37.1	44.1	50.7	52.5	57.7	48.7	52.1	56.1	59.0	60.5
MPSR+Aug	49.9	56.2	61.5	63.0	66.5	37.4	43.0	51.4	53.6	58.6	48.1	49.3	55.7	58.7	60.8
MPSR+Meta-Static+Aug	51.3	56.9	62.0	62.8	66.9	37.7	43.5	50.7	53.7	58.1	48.6	49.5	55.9	58.5	60.3
MPSR+Meta-Dynamic+Aug	51.3	56.8	62.1	63.3	67.0	37.8	44.2	51.7	54.3	59.3	48.9	50.5	56.5	59.0	61.2
MPSR+Meta-ScaledDynamic+Aug	51.9	57.6	62.4	63.7	67.6	37.8	44.9	51.9	54.2	59.4	49.2	51.9	56.7	59.7	61.1
DeFRCN+Meta-ScaledDynamic+Aug	**66.7**	**69.3**	**69.8**	**72.2**	**72.1**	**47.7**	**55.8**	**61.8**	**63.9**	**63.7**	**64.9**	**65.8**	**66.2**	**69.7**	**70.2**

Table 7.9: Comparison to meta-learning based FSOD methods on the Pascal VOC dataset, with only novel classes.

Method/Shot		Novel Set 1					Novel Set 2					Novel Set 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
ML	M. R-CNN [64] (ICCV'19)	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
	M. R-CNN* [64] (ICCV'19)	16.8	20.1	20.3	38.2	43.7	7.7	12.0	14.9	21.9	31.1	9.2	13.9	26.2	29.2	36.2
	FSRW [10] (ICCV'19)	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	39.2	19.2	21.7	25.7	40.6	41.3
	MetaDet [240] (ICCV'19)	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
	FsDet [60] (ECCV'20)	25.4	20.4	37.4	36.1	42.3	22.9	21.7	22.6	25.6	29.2	32.4	19.0	29.8	33.2	39.8
	TIP [194] (CVPR'21)	27.7	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9
	DCNet [192] (CVPR'21)	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
	CME [62] (CVPR'21)	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
	QA-FewDet [193] (ICCV'21)	41.0	33.2	35.3	47.5	52.0	23.5	29.4	37.9	35.9	37.1	33.2	29.4	37.6	39.8	41.5
	KFSOD [67] (CVPR'22)	44.6	-	54.4	60.9	65.8	**37.8**	-	43.1	48.1	50.4	34.8	-	44.1	52.7	53.9
	FCT [68] (CVPR'22)	**49.9**	**57.1**	**57.9**	**63.2**	**67.1**	27.6	**34.5**	**43.7**	**49.2**	**51.2**	**39.5**	**54.7**	**52.3**	**57.0**	**58.7**
Ours	DeFRCN+Meta-ScaledDynamic+Aug	**58.4**	**62.4**	**63.2**	**67.6**	**67.7**	**34.0**	**43.1**	**51.0**	**53.6**	**54.0**	**55.1**	**56.6**	**57.3**	**62.6**	**63.7**

Table 7.10: Comparison to meta-learning G-FSOD methods on the Pascal VOC dataset, with both base and novel classes.

Method/Shot		Split-1					Split-2					Split-3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
ML	M. R-CNN [64] (ICCV'19)	17.3	25.3	27.3	44.4	50.4	11.6	18.5	21.9	30.8	41.3	13.3	20.2	33.4	38.0	45.5
	FSRW [10] (ICCV'19)	24.2	24.8	37.8	44.2	54.2	25.5	24.9	33.8	41.5	49.0	29.7	32.4	36.7	49.9	49.9
	FsDet [60] (ECCV'20)	31.1	28.4	39.1	43.5	49.5	29.3	30.5	30.7	34.4	39.8	35.2	26.9	35.6	41.8	47.8
Ours	DeFRCN+Meta-ScaledDynamic+Aug	**58.4**	**62.4**	**63.2**	**67.6**	**67.7**	**34.0**	**43.1**	**51.0**	**53.6**	**54.0**	**55.1**	**56.6**	**57.3**	**62.6**	**63.7**

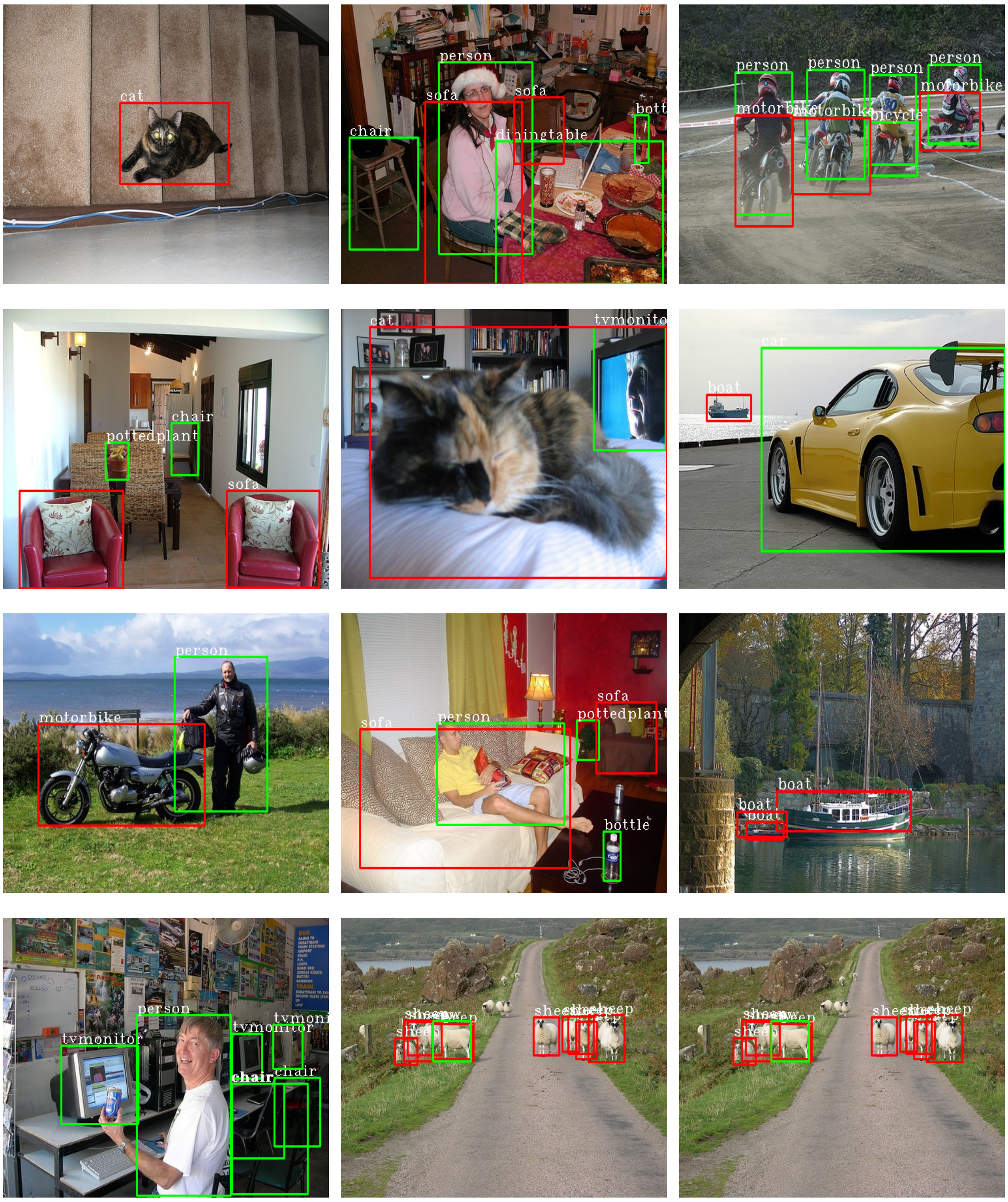


Figure 7.5: Randomly sampled *MPSR+Meta-ScaledDynamic* object detection results for the Pascal VOC dataset Split-3/10-shot experiment.

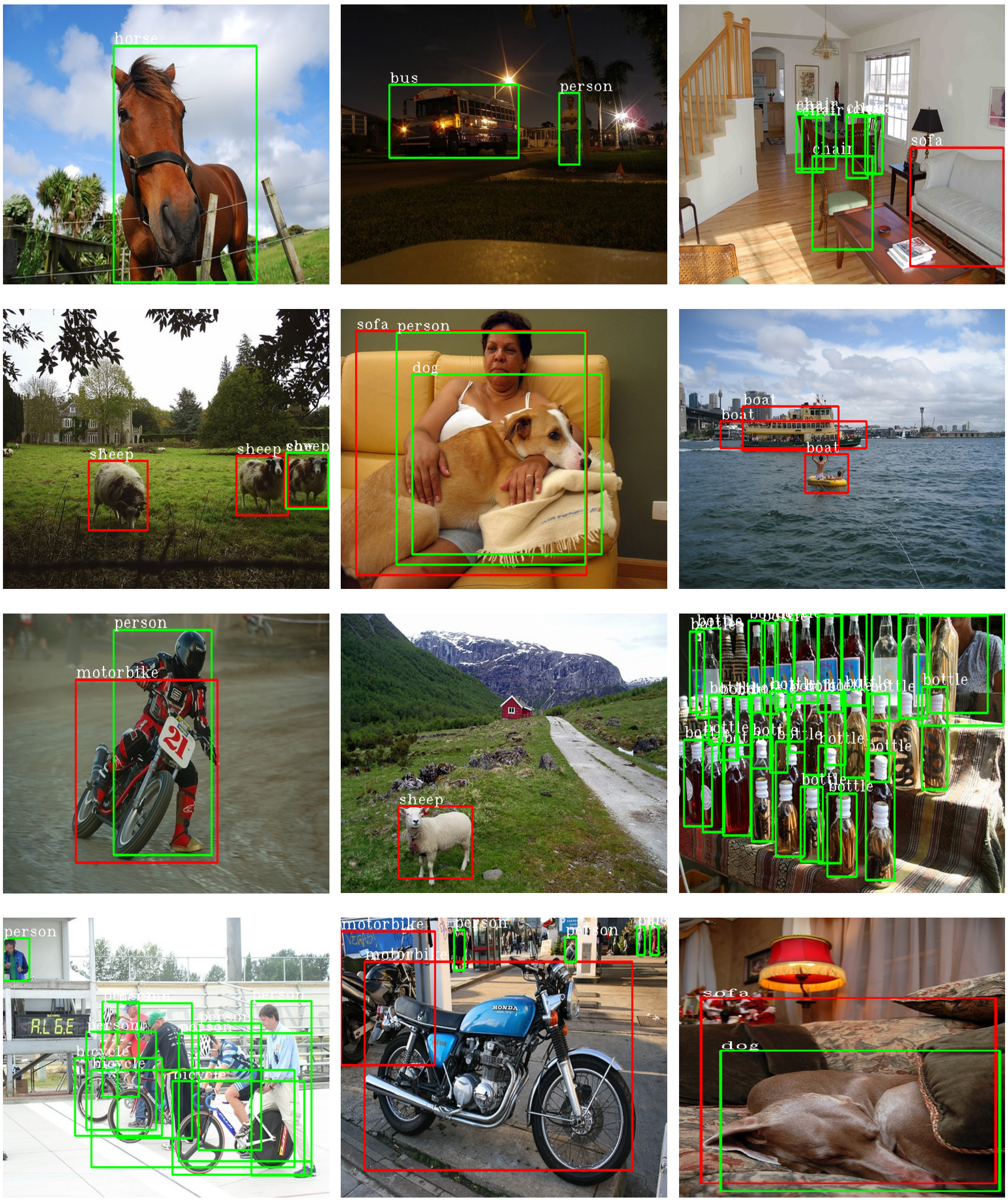


Figure 7.6: Randomly sampled *MPSR+Meta-ScaledDynamic+Aug* object detection results for the Pascal VOC dataset Split-3/10-shot experiment.

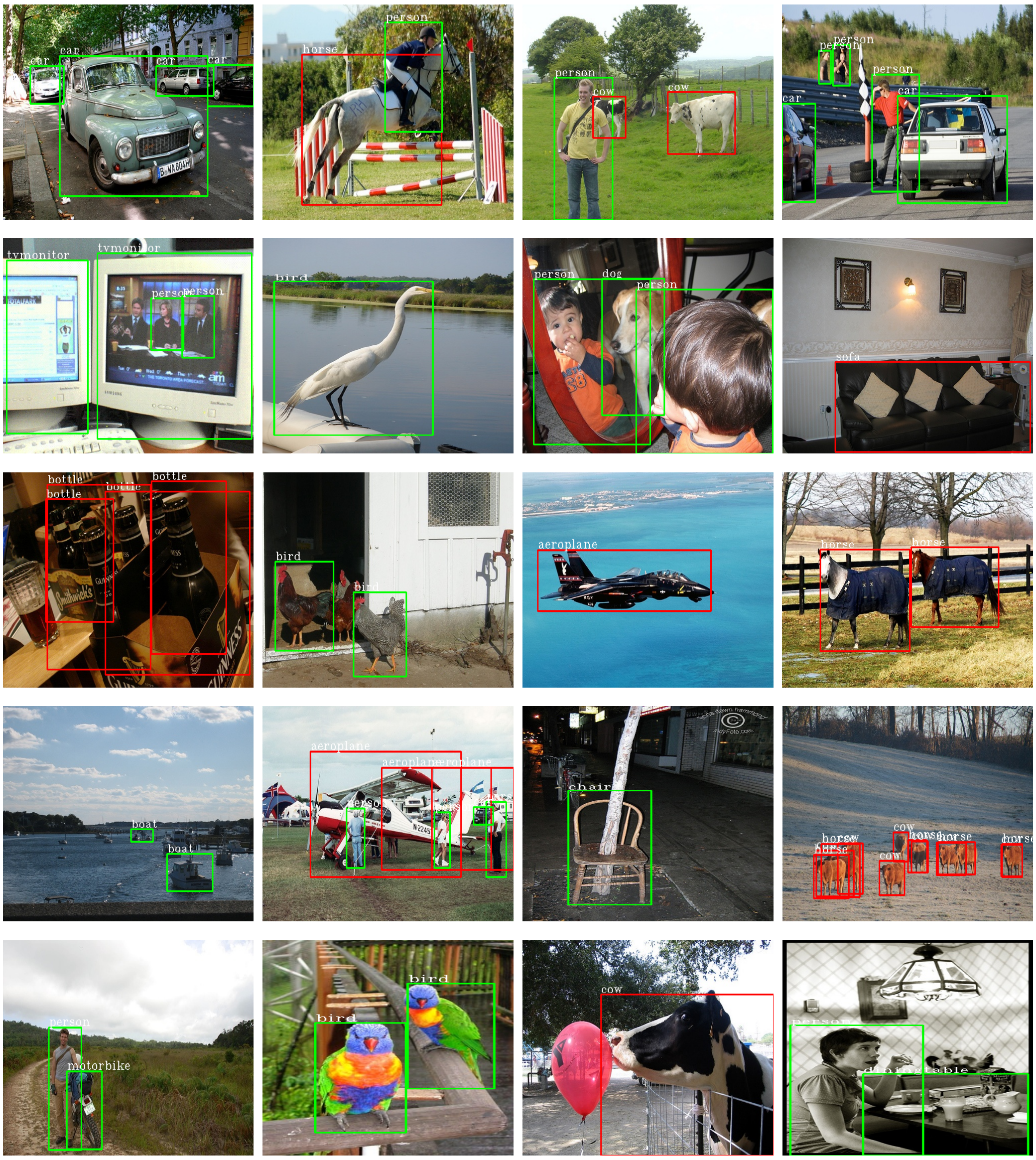


Figure 7.7: Randomly sampled *DeFRCN+Meta-ScaledDynamic+Aug* object detection results for the Pascal VOC dataset Split-2/10-shot experiment.

Table 7.11: FSOD and G-FSOD results on the MS COCO dataset with novel classes.

Method/Shot		Novel Classes		All Classes (HM)	
		10-shot	30-shot	10-shot	30-shot
ML	ONCE [65]	1.2	-	2.2	-
	Meta R-CNN [64]	6.1	9.9	5.6	8.3
	FSRW [10]	5.6	9.1	-	-
	FsDetView [239]	7.6	12.0	6.9	10.5
	TIP [194]	16.3	18.3	-	-
	DCNET [241]	12.8	18.6	-	-
	CME [62]	15.1	16.9	-	-
	QA-FewDet [193]	10.2	11.5	-	-
	FCT [68]	17.1	21.4	-	-
Ours	DeFRCN+Meta-ScaledDynamic+Aug	**18.8**	**23.4**	**24.4**	**28.0**

CHAPTER 8

CONCLUSIONS

In this thesis, we studied approaches to handle the object detection problem with minimal supervision. These approaches include the ZSD problem that we propose in the thesis and the FSOD problem being studied in the literature. Our aim in defining the ZSD problem is to adapt the ZSL concepts to different scenarios, because localizing the unseen class instances is as important as recognition in various applications, such as robotics. Besides, when we consider the long-tail distribution problem within the scope of object detection, many labeled classes are necessary, and making annotations for them is laborious.

Our proposed first ZSD approach aggregates both label embeddings and convex combinations of semantic embeddings together in a region embedding framework. We also prepared two different datasets to analyze the proposed ZSD approach: the first of these is the Fashion-ZSD, which is the toy dataset we generated from the Fashion-MNIST dataset, and the other one is Pascal VOC dataset, whose ZSD splits we set in the literature. Experimental results showed that our proposed problem and novel approach obtained promising results. After defining the first ZSD approach, our studies for the ZSD problem are on background modeling for unseen classes, label embedding techniques that are frequently used in ZSD models, and the definition of the ZSIC problem as an extension of ZSD.

One of the most important shortcomings of the current ZSD methods is the presence of unlabeled instances of unseen classes in the images during the training time. In such a scenario, due to the nature of the working mechanism of existing object detection models, negative region samples might be collected from unseen class regions and samples belonging to these classes might be modeled as background. In this case, although the proposed approaches are plausible, object detection models may not generate candidate regions for these classes during inference. As a first attempt at this problem, we provide a textual attention mechanism to ZSD models so that pre-RPN features become to be class-specific, and candidate regions belonging to unseen classes can be generated even if instances of these classes exist in the training set. The ZSD approaches use knowledge transfer from seen classes to unseen classes, as in other ZSL models. For this reason, we also analyzed label embedding concepts within the scope of the thesis. We have built embedding models to make the information transfer more accurate and visually more meaningful.

In this thesis, we also defined the *true zero-shot image captioning* problem as a continuation of our proposed approach for the ZSD problem. An important shortcoming of current image captioning methods (*i.e. partial zero-shot image captioning*) that aim training through non-paired datasets is that they do not work in a fully ZSL setting. These methods generate captions for images that consist of classes not seen in captioning datasets, but they assume that there is a ready-to-use fully supervised visual recognition model, so they are not complete zero-shot learning approaches. Alternatively, we proposed a GZSD model that uses a novel practical class embedding scheme and class scaling instead of the aforementioned ready-to-use fully supervised object detection models. Also, we defined a metric we named V-METEOR by considering current evaluation metrics are not sufficient for the ZSIC problem.

Another concept that we are interested in in this thesis is the FSOD problem, in which fine-tuning and meta-learning-based approaches are proposed. Fine-tuning based methods provide simple and reliable approaches by focusing on the handcrafted design of fine-tuning details for few-shot training. For this reason, we provide automatic learning of the parameters of loss functions or augmentation magnitudes in fine-tuning based methods with an intermediate learning step we call meta-tuning. The proposed tuning scheme uses meta-learning principles with reinforcement learning, and obtains interpretable loss functions and augmentation magnitudes for few-shot training. Experimental results with our meta-tuning method, which we built on various baseline models, show that the proposed idea obtains state-of-the-art results in benchmark Pascal VOC and MS COCO datasets.

As a future research direction, the approaches to be proposed within the scope of ZSD should be in more realistic scenarios. One of the important steps to be taken in this regard is background modeling, which we mentioned in the thesis and proposed as the first approach. Otherwise, a dilemma occurs as we mentioned in the previous chapters. This situation has recently been expressed for the FSOD problem [74]. Moreover, in order to increase the performance, the use of generative models [3, 93, 94, 95, 96, 97, 98, 99, 100, 101], in which ZSL methods have recently evolved, will be important for ZSD problem.

In this thesis, we also propose the novel meta-tuning approach for the FSOD problem. This approach allows learning inductive biases that can boost FSOD by applying meta-learning principles to fine-tuning based methods. We use the proposed meta-tuning approach for classification loss terms and augmentation magnitudes, but this idea is applicable to many different parameters. As another future research direction, meta-tuning or derived ideas can be used from a broader perspective. For example, some layers of models can also be learned through this RL-based approach. As another future research direction, a meta-tuning approach can also be proposed that will use the mAP scores learned during model training more efficiently. In the current model, mAP results are used and discarded as instant rewards, these mAP values can also be stored for following learning steps.

REFERENCES

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 819–826, IEEE, 2013.

[2] M. Elhoseiny, B. Saleh, and A. Elgammal, "Write a classifier: Zero-shot learning using purely textual descriptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2584–2591, 2013.

[3] M. Bucher, S. Herbin, and F. Jurie, "Generating visual representations for zero-shot classification," in *Proc. IEEE Int. Conf. on Computer Vision Workshops*, pp. 2666–2673, 2017.

[4] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized Zero-Shot Learning with Deep Calibration Network," in *NeurIPS*, pp. 2005–2015, 2018.

[5] S. Rahman, S. Khan, and F. Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," *arXiv preprint arXiv:1803.06049*, 2018.

[6] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko, "Detector discovery in the wild: Joint multiple instance and representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2883–2891, 2015.

[7] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574, 2016.

[8] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9962–9971, 2020.

[9] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to Learn Quickly for Few-Shot Learning," *arXiv e-prints*, p. arXiv:1707.09835, July 2017.

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

[11] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," *arXiv preprint arXiv:2003.06957*, 2020.

[12] P. Liu, G. Zhang, B. Wang, H. Xu, X. Liang, Y. Jiang, and Z. Li, "Loss function discovery for object detection via convergence-simulation driven search," *arXiv preprint arXiv:2102.04700*, 2021.

[13] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[15] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, 2017.

[16] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[18] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proceedings of International Conference on Learning Representations*, 2014.

[19] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2874–2883, 2016.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[22] J. Yan, Z. Lei, L. Wen, and S. Li, “The fastest deformable part model for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2497–2504, 2014.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[24] H. Law and J. Deng, “Cornernet: Detecting objects as paired keypoints,” in *European conference on computer vision*, pp. 734–750, 2018.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[26] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, Ieee, 2001.

[27] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1, pp. 886–893, Ieee, 2005.

[28] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Ieee, 2008.

[29] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” *arXiv preprint arXiv:2207.02696*, 2022.

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of

deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[32] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

[37] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[38] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*, pp. 4055–4064, PMLR, 2018.

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

[40] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10213–10224, 2021.

[41] Z. Huang, Y. Zou, B. Kumar, and D. Huang, "Comprehensive attention self-

distillation for weakly-supervised object detection," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[42] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[43] Y. Liu, Z. Zhang, L. Niu, J. Chen, and L. Zhang, "Mixed supervised object detection by transferring mask prior and semantic similarity," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.

[44] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *Proceedings of International Conference on Learning Representations*, 2014.

[45] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1641–1648, IEEE, 2011.

[46] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, "Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct 2017.

[47] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1425–1438, 2015.

[48] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015.

[49] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine learning*, vol. 81, no. 1, pp. 21–35, 2010.

[50] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[51] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[52] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent embeddings for zero-shot classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 69–77, 2016.

[53] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, “Evaluation of output embeddings for fine-grained image classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2015.

[54] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen, “Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5975–5984, 2016.

[55] S. Rahman, S. Khan, and N. Barnes, “Improved visual-semantic alignment for zero-shot object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11932–11939, 2020.

[56] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, “Semantics-guided contrastive network for zero-shot object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. (early access), 2022.

[57] J. Lu, J. Yang, D. Batra, and D. Parikh, “Neural baby talk,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228, 2018.

[58] T. Yao, Y. Pan, Y. Li, and T. Mei, “Incorporating copying mechanism in image captioning for learning novel objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6580–6588, 2017.

[59] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl*

workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72, 2005.

[60] Y. Xiao and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," in *European conference on computer vision*, pp. 192–210, Springer, 2020.

[61] G. Zhang, Z. Luo, K. Cui, and S. Lu, "Meta-detr: Few-shot object detection via unified image-level meta-learning," *arXiv preprint arXiv:2103.11731*, vol. 2, 2021.

[62] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7363–7372, 2021.

[63] T.-I. Chen, Y.-C. Liu, H.-T. Su, Y.-C. Chang, Y.-H. Lin, J.-F. Yeh, W.-C. Chen, and W. Hsu, "Dual-awareness attention for few-shot object detection," *IEEE Transactions on Multimedia*, 2021.

[64] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9577–9586, 2019.

[65] J.-M. Perez-Rua, X. Zhu, T. M. Hospedales, and T. Xiang, "Incremental few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13846–13855, 2020.

[66] L. Yin, J. M. Perez-Rua, and K. J. Liang, "Sylph: A hypernetwork framework for incremental few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9035–9045, 2022.

[67] S. Zhang, L. Wang, N. Murray, and P. Koniusz, "Kernelized few-shot object detection with efficient integral aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19207–19216, 2022.

[68] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5321–5330, 2022.

[69] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection

via contrastive proposal encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7352–7362, 2021.

[70] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *European conference on computer vision*, pp. 456–472, Springer, 2020.

[71] Z. Fan, Y. Ma, Z. Li, and J. Sun, "Generalized few-shot object detection without forgetting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4527–4536, 2021.

[72] W. Zhang and Y.-X. Wang, "Hallucination improves few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13008–13017, 2021.

[73] Y. Cao, J. Wang, Y. Jin, T. Wu, K. Chen, Z. Liu, and D. Lin, "Few-shot object detection via association and discrimination," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16570–16581, 2021.

[74] P. Kaul, W. Xie, and A. Zisserman, "Label, verify, correct: A simple few shot object detection method," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14237–14247, 2022.

[75] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "Defrcn: Decoupled faster r-cnn for few-shot object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8681–8690, 2021.

[76] T. Jeong and H. Kim, "Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 3907–3916, 2020.

[77] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017.

[78] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[79] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[80] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 951–958, 2009.

[81] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, “Synthesized classifiers for zero-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[82] E. Kodirov, T. Xiang, and S. Gong, “Semantic autoencoder for zero-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3174–3183, 2017.

[83] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam, “Large-scale object classification using label relation graphs,” in *European conference on computer vision*, pp. 48–64, Springer, 2014.

[84] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, *et al.*, “Devise: A deep visual-semantic embedding model,” in *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

[85] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, “Transductive unbiased embedding for zero-shot learning,” *arXiv preprint arXiv:1803.11320*, 2018.

[86] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, “Zero-shot learning by convex combination of semantic embeddings,” *arXiv preprint arXiv:1312.5650*, 2013.

[87] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li, “Zero-shot learning using synthesised unseen visual data with diffusion regularisation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, 2018.

[88] C. Luo, Z. Li, K. Huang, J. Feng, and M. Wang, “Zero-shot learning via attribute regression and class prototype rectification,” *IEEE Trans. on Image Processing*, vol. 27, no. 2, pp. 637–648, 2018.

[89] Y. Yu, Z. Ji, J. Guo, and Z. Zhang, “Zero-shot learning via latent space encoding,” *IEEE transactions on cybernetics*, no. 99, pp. 1–12, 2018.

[90] Y. Feng, X. Huang, P. Yang, J. Yu, and J. Sang, “Non-generative generalized zero-shot learning via task-correlated disentanglement and controllable samples synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9346–9355, 2022.

[91] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," vol. 41, no. 9, pp. 2251–2265, 2018.

[92] J. Chen, Y. Geng, Z. Chen, I. Horrocks, J. Z. Pan, and H. Chen, "Knowledge-aware zero-shot learning: Survey and perspective," *arXiv preprint arXiv:2103.00070*, 2021.

[93] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *European conference on computer vision*, 2018.

[94] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, pp. 2188–2196, 2018.

[95] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10275–10284, 2019.

[96] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, "A generative adversarial approach for zero-shot learning from noisy texts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1004–1013, 2018.

[97] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7402–7411, 2019.

[98] M. B. Sariyildiz and R. G. Cinbis, "Gradient matching generative networks for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2168–2178, 2019.

[99] S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, and L. Shao, "Free: Feature refinement for generalized zero-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 122–131, 2021.

[100] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8712–8720, 2021.

[101] H. Su, J. Li, Z. Chen, L. Zhu, and K. Lu, "Distinguishing unseen from seen for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7885–7894, 2022.

[102] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of International Conference on Learning Representations*, 2014.

[103] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of International Conference on Machine Learning*, pp. 1278–1286, PMLR, 2014.

[104] M. Elhoseiny and M. Elfeki, "Creativity inspired zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5784–5793, 2019.

[105] Y. H. Li, T.-Y. Chao, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Make an omelette with breaking eggs: Zero-shot learning for novel attribute synthesis," *Advances in Neural Information Processing Systems*, 2022.

[106] H. Jiang, R. Wang, S. Shan, and X. Chen, "Transferable Contrastive Network for Generalized Zero-Shot Learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Aug. 2019.

[107] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *European conference on computer vision*, pp. 52–68, Springer, 2016.

[108] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "Dsod: Learning deeply supervised object detectors from scratch," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1919–1927, 2017.

[109] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[110] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[111] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, pp. 379–387, 2016.

[112] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," *arXiv preprint arXiv:1804.04340*, 2018.

[113] S. Rahman, S. Khan, and N. Barnes, "Polarity loss for zero-shot object detection," *arXiv preprint arXiv:1811.08982*, 2018.

[114] S. Rahman, S. Khan, and N. Barnes, "Transductive learning for zero-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6082–6091, 2019.

[115] Z. Li, L. Yao, X. Zhang, X. Wang, S. Kanhere, and H. Zhang, "Zero-shot object detection with textual descriptions," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 33, pp. 8690–8697, 2019.

[116] Y. Shao, Y. Li, and D. Wang, "Zero-shot detection with transferable object proposal mechanism," in *IEEE Int. Conf. on Image Processing*, pp. 3666–3670, IEEE, 2019.

[117] D. Gupta, A. Anantharaman, N. Mamgain, V. N. Balasubramanian, C. Jawahar, *et al.*, "A multi-space approach to zero-shot object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1209–1217, 2020.

[118] Y. Li, P. Li, H. Cui, and D. Wang, "Inference fusion with associative semantics for unseen object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1993–2001, 2021.

[119] Y. Zheng, R. Huang, C. Han, X. Huang, and L. Cui, "Background learnable cascade for zero-shot object detection," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[120] Y. Zheng, X. Huang, and L. Cui, "Visual language based succinct zero-shot object detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5410–5418, 2021.

[121] H. Nie, R. Wang, and X. Chen, "From node to graph: Joint reasoning on visual-semantic relational graph for zero-shot detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1109–1118, 2022.

[122] P. Huang, J. Han, D. Cheng, and D. Zhang, "Robust region feature synthesizer

for zero-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7622–7631, 2022.

[123] C. Yang, W. Wu, Y. Wang, and H. Zhou, "A novel feature-based model for zero-shot object detection with simulated attributes," *Applied Intelligence*, vol. 52, no. 6, pp. 6905–6914, 2022.

[124] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*, pp. 391–405, 2014.

[125] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," in *Advances in Neural Information Processing Systems*, pp. 2005–2015, 2018.

[126] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko, "Lsda: Large scale detection through adaptation," in *Advances in Neural Information Processing Systems*, pp. 3536–3544, 2014.

[127] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, 2016.

[128] A. Arun, C. Jawahar, and M. P. Kumar, "Dissimilarity Coefficient Based Weakly Supervised Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Long Beach, CA, USA), pp. 9424–9433, June 2019.

[129] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, Y. J. Lee, A. G. Schwing, and J. Kautz, "Instance-aware, context-focused, and memory-efficient weakly supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10598–10607, 2020.

[130] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *Proceedings of International Conference on Learning Representations*, 2015.

[131] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659, 2016.

[132] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual

attention," in *Proceedings of International Conference on Machine Learning*, pp. 2048–2057, 2015.

[133] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[134] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

[135] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, 2013.

[136] Y. Hirota, Y. Nakashima, and N. Garcia, "Quantifying societal bias amplification in image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13450–13459, 2022.

[137] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao, and J. Zhao, "Show, deconfound and tell: Image captioning with causal inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18041–18050, 2022.

[138] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17980–17989, 2022.

[139] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18030–18040, 2022.

[140] Y. Li, Y. Pan, T. Yao, and T. Mei, "Comprehending and ordering semantics for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17990–17999, 2022.

[141] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*, pp. 15–29, Springer, 2010.

[142] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[143] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in *Advances in Neural Information Processing Systems*, pp. 1143–1151, 2011.

[144] C. Sun, C. Gan, and R. Nevatia, “Automatic concept discovery from parallel text and visual corpora,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2596–2604, 2015.

[145] L. Yang, K. Tang, J. Yang, and L.-J. Li, “Dense captioning with joint inference and visual context,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2193–2202, 2017.

[146] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, “Dense-captioning events in videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 706–715, 2017.

[147] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2016.

[148] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, “Captioning images with diverse objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5753–5761, 2017.

[149] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Guided open vocabulary image captioning with constrained beam search,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 936–945, 2017.

[150] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, “Decoupled novel object captioner,” in *Proc. of the 26th ACM international conf. on Multimedia*, pp. 1029–1037, 2018.

[151] R. Khan, M. S. Islam, K. Kanwal, M. Iqbal, M. Hossain, Z. Ye, *et al.*, “A deep neural framework for image caption generation using gru-based attention mechanism,” *arXiv preprint arXiv:2203.01594*, 2022.

[152] A. Yuan, X. Li, and X. Lu, “3g structure for image caption generation,” *Neurocomputing*, vol. 330, pp. 17–28, 2019.

[153] L. Cheng, W. Wei, X. Mao, Y. Liu, and C. Miao, “Stack-vs: Stacked visual-semantic attention for image caption generation,” *IEEE Access*, vol. 8, pp. 154953–154965, 2020.

[154] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[155] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in *European conference on computer vision*, pp. 382–398, Springer, 2016.

[156] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[157] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[158] Z. Wang, B. Feng, K. Narasimhan, and O. Russakovsky, “Towards unique and informative captioning of images,” in *European conference on computer vision*, pp. 629–644, 2020.

[159] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proceedings of International Conference on Machine Learning*, vol. 70, pp. 1126–1135, 2017.

[160] A. Nichol and J. Schulman, “Reptile: a scalable metalearning algorithm,” *arXiv: Learning*, 2018.

[161] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, “Meta-learning with implicit gradients,” in *NeurIPS*, 2019.

[162] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, “Meta-learning with latent embedding optimization,” in *Proceedings of International Conference on Learning Representations*, 2019.

[163] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler, “Rapid adaptation with conditionally shifted neurons,” in *ICML*, 2018.

[164] E. Park and J. B. Oliva, “Meta-curvature,” in *NeurIPS*, 2019.

[165] K. Cao, M. Brbić, and J. Leskovec, “Concept learners for few-shot learning,” in *Proceedings of International Conference on Learning Representations*, 2021.

[166] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. Hwang, and Y. Yang, “Learning to propagate labels: Transductive propagation network for few-shot learning,” in *Proceedings of International Conference on Learning Representations*, 2019.

[167] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[168] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

[169] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in *Proceedings of International Conference on Machine Learning*, vol. 48, pp. 1842–1850, 2016.

[170] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[171] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, “Few-shot learning via embedding adaptation with set-to-set functions,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8805–8814, 2020.

[172] B. N. Oreshkin, P. R. Lopez, and A. Lacoste, “Tadam: Task dependent adaptive metric for improved few-shot learning,” in *NeurIPS*, 2018.

[173] C. Zhang, Y. Cai, G. Lin, and C. Shen, “Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[174] H. Yao, L. Zhang, and C. Finn, “Meta-learning with fewer tasks through task interpolation,” in *Proceeding of the 10th International Conference on Learning Representations*, 2022.

[175] B. Hariharan and R. B. Girshick, “Low-shot visual recognition by shrinking and

hallucinating features," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3037–3046, 2017.

[176] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-Shot Learning from Imaginary Data," *arXiv:1801.05401 [cs]*, Jan. 2018.

[177] M. Lazarou, Y. Avrithis, and T. Stathaki, "Tensor feature hallucination for few-shot learning," *ArXiv*, vol. abs/2106.05321, 2021.

[178] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need?," in *European conference on computer vision*, 2020.

[179] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[180] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv preprint arXiv:1911.04623*, 2019.

[181] S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. M. Hospedales, "Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference," *arXiv e-prints*, p. arXiv:2204.07305, Apr. 2022.

[182] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, "Improved Few-Shot Visual Classification," *arXiv e-prints*, p. arXiv:1912.03432, Dec. 2019.

[183] P. Bateni, J. Barber, J.-W. van de Meent, and F. Wood, "Enhancing few-shot image classification with unlabelled examples," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2796–2805, January 2022.

[184] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," *arXiv preprint arXiv:2003.12060*, 2020.

[185] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *Proceedings of International Conference on Learning Representations*, 2019.

[186] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10649–10657, 2019.

[187] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A Baseline for Few-Shot Image Classification," in *ICLR*, p. 20, 2020.

[188] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A Closer Look at Few-shot Classification," in *ICLR 2019*, 2019.

[189] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4013–4022, 2020.

[190] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," *arXiv preprint arXiv:1911.12529*, 2019.

[191] L. Zhang, S. Zhou, J. Guan, and J. Zhang, "Accurate few-shot object detection with support-query mutual guidance and hybrid loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14424–14432, 2021.

[192] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10185–10194, 2021.

[193] G. Han, Y. He, S. Huang, J. Ma, and S.-F. Chang, "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3263–3272, 2021.

[194] A. Li and Z. Li, "Transformation invariant few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3094–3102, 2021.

[195] A. Wu, S. Zhao, C. Deng, and W. Liu, "Generalized and discriminative few-shot object detection via svd-dictionary enhancement," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6353–6364, 2021.

[196] E. Real, C. Liang, D. So, and Q. Le, "Automl-zero: Evolving machine learn-

ing algorithms from scratch," in *Proceedings of International Conference on Machine Learning*, pp. 8007–8019, PMLR, 2020.

[197] S. Gonzalez and R. Miikkulainen, "Improved training speed, accuracy, and data utilization through loss function optimization," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, IEEE, 2020.

[198] C. Li, X. Yuan, C. Lin, M. Guo, W. Wu, J. Yan, and W. Ouyang, "Am-lfs: Automl for loss function search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8410–8419, 2019.

[199] X. Wang, S. Wang, C. Chi, S. Zhang, and T. Mei, "Loss function search for face recognition," in *Proceedings of International Conference on Machine Learning*, pp. 10029–10038, PMLR, 2020.

[200] D. Ho, E. Liang, X. Chen, I. Stoica, and P. Abbeel, "Population based augmentation: Efficient learning of augmentation policy schedules," in *Proceedings of International Conference on Machine Learning*, pp. 2731–2741, PMLR, 2019.

[201] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[202] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, pp. 702–703, 2020.

[203] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal on Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[204] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1778–1785, IEEE, 2009.

[205] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009.

[206] J. Sánchez and M. Molina, "Trading-off information modalities in zero-shot classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3841–3849, 2022.

[207] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, "Vgse: Visually-grounded semantic embeddings for zero-shot learning," *arXiv preprint arXiv:2203.10444*, 2022.

[208] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2371–2381, 2021.

[209] Y. Zheng, J. Wu, Y. Qin, F. Zhang, and L. Cui, "Zero-shot instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2593–2602, 2021.

[210] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of koschmieder's model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16205–16215, 2021.

[211] Z. Cheng, Z. Xiong, C. Chen, D. Liu, and Z.-J. Zha, "Light field super-resolution with zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10010–10019, 2021.

[212] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu, "Rethinking classification and localization for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10186–10195, 2020.

[213] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4282–4291, 2019.

[214] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[215] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[216] J. Lei Ba, K. Swersky, S. Fidler, and R. Salakhutdinov, "Predicting deep zero-shot convolutional neural networks using textual descriptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015.

[217] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[218] Z. Al-Halah and R. Stiefelhagen, "How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pp. 837–843, IEEE, 2015.

[219] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, "Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning," in *ICCV*, 2017.

[220] C. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, pp. 453–465, March 2014.

[221] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results." http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[222] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations*, 2015.

[223] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[224] H. Zhang, Y. Long, L. Liu, and L. Shao, "Adversarial unseen visual feature synthesis for zero-shot learning," *Neurocomputing*, vol. 329, 2019.

[225] L. Bo and C. Sminchisescu, "Twin gaussian processes for structured prediction," *International Journal on Computer Vision*, vol. 87, no. 1-2, pp. 28–52, 2010.

[226] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," *arXiv preprint arXiv:1509.04767*, 2015.

[227] X. Yin and V. Ordonez, "Obj2text: Generating visually descriptive language from object layouts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 177–187, 2017.

[228] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Advances in Neural Information Processing Systems*, pp. 2692–2700, 2015.

[229] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, “Re-evaluating automatic metrics for image captioning,” in *Proc. of the 15th Conference of the European Chapter of the Assoc. for Computational Linguistics*, 2017.

[230] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[231] D. Hoiem, Y. Chodpathumwan, and Q. Dai, “Diagnosing error in object detectors,” in *European conference on computer vision*, pp. 340–353, 2012.

[232] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, “Simple copy-paste is a strong data augmentation method for instance segmentation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2021.

[233] L. Chen, T. Yang, X. Zhang, W. Zhang, and J. Sun, “Points as queries: Weakly semi-supervised object detection by points,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8819–8828, 2021.

[234] X. Wu, D. Sahoo, and S. Hoi, “Meta-rcnn: Meta learning for few-shot object detection,” in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1679–1687, 2020.

[235] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. B. Ayed, and J. Dolz, “Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need?,” in *arXiv:2012.06166 [cs]*, 2021.

[236] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[237] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[238] M. Papini, A. Battistello, and M. Restelli, “Balancing learning speed and stability in policy gradient via adaptive exploration,” in *Proc. Int. Conf. on Artif. Intellig. and Stat.*, pp. 1188–1199, 2020.

[239] Y. Xian, B. Schiele, and Z. Akata, “Zero-shot learning-the good, the bad and the ugly,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4582–4591, 2017.

[240] Y.-X. Wang, D. Ramanan, and M. Hebert, “Meta-learning to detect rare objects,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9925–9934, 2019.

[241] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “Centernet: Keypoint triplets for object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6569–6578, 2019.

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname: Berkan Demirel
Nationality: Turkish

EDUCATION

- MS, Dept. of Computer Engineering, Hacettepe University, 2016
- BS, Dept. of Computer Engineering, Hacettepe University, 2014

PROFESSIONAL EXPERIENCE

- HAVELSAN Inc. (2015 - today)
- MilSOFT Inc. (2013 - 2015)

PUBLICATIONS

- **Berkan Demirel**, Orhun Buğra Baran, Ramazan Gokberk Cinbis, "Meta-tuning Loss Functions and Data Augmentation for Few-shot Object Detection". **(Submitted)**.

1. **Berkan Demirel**, Orkun Öztürk, Mehmet Can Baytekin, Ramazan Gokberk Cinbis, "Zero-shot Object Detection in the Wild". **(Submitted)**

- **Berkan Demirel**, Ramazan Gokberk Cinbis, "Caption Generation on Scenes with Seen and Unseen Object Categories", Image and Vision Computing (IMAVIS), 2022.

- A.O. Tur, B. Selbes, H.I. Ozturk, I. Karakaya, **B. Demirel**, "Importance of Image Enhancement Methods for Fingerprint Recognition", 29th Signal Processing and Communications Applications Conference (SIU), 2021.
- A. O. Cimtay, B. Alkan, **B. Demirel**, "Fingerprint Pattern Classification by Using Various Pre-Trained Deep Neural Networks", 2nd International Conference On Access To Recent Advances In Engineering And Digitalization (ARACONF), 2021.
- I. Karakaya, **B. Demirel**, O. Öztürk, M. Bal, E. Başeski, "HVLSeg: An Ensemble Instance Segmentation Model on Satellite Images", 28th Signal Processing and Communications Applications Conference (SIU), 2020.
- **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Image Captioning with Unseen Objects", British Machine Vision Conference (BMVC), September 2019.
- **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Learning Visually Consistent Label Embeddings for Zero-Shot Learning", IEEE International Conference on Image Processing (ICIP), September 2019.
- Omer Ozdil, Yunus Emre Esin, **Berkan Demirel**, Safak Ozturk, "Generating Spectral Signature Library for Patterned Object in Hyperspectral Images", 2019 9th International Conference on Recent Advances in Space Technologies (RAST), June 2019.
- O. Ozdil, A. Gunes, Y.E. Esin, **Berkan Demirel**, S. Ozturk, "Comparison of Target Detection Performance for Radiance and Reflectance Domain in VNIR Hyperspectral Images", 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 2019.
- **Berkan Demirel**, Yunus Emre Esin, Ömer Özdil, and Safak Ozturk, "Segmentation-Aware Hyperspectral Image Classification", 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 2019.
- Yunus Emre Esin, **Berkan Demirel**, Omer Ozdil, Safak Ozturk, "Ortho-Rectification of Hyperspectral Camera Data with Central Processing Unit and Graphics Processing Unit", 2019 9th International Conference on Recent Advances in Space Technologies (RAST), June 2019.

- Yunus Emre Esin, Omer Ozdil, **Berkan Demirel**, Safak Ozturk, "Practical Focus Adjustment Method for Hyperspectral Cameras”, 2019 9th International Conference on Recent Advances in Space Technologies (RAST), June 2019.
- **Berkan Demirel**, Yunus Emre Esin, Ömer Özdil, and Safak Ozturk. "Hyperspectral Target Detection Using Long Short-Term Memory and Spectral Angle Mapper.” Signal Processing and Communications Applications Conference (SIU), 2019 27th. IEEE, 2019.
- **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Zero-Shot Object Detection by Hybrid Region Embedding”, British Machine Vision Conference (BMVC), August 2018.
- Omer Ozdil, **Berkan Demirel**, Yunus Emre Esin, and Safak Ozturk. "SPARK detection with thermal camera.” Signal Processing and Communications Applications Conference (SIU), 2018 26th. IEEE, 2018.
- Safak Ozturk, Yunus Emre Esin, Yusuf Artan, Omer Ozdil, **Berkan Demirel**, "Importance of Band Selection for Ethene and Methanol Gas Detection in Hyperspectral Imagery.” 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). IEEE, 2018.
- Omer Ozdil, Ahmet Gunes, Yunus Emre Esin, Safak Ozturk, **Berkan Demirel**, "4-Stage Target Detection Approach In Hyperspectral Images.” 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). IEEE, 2018.
- Omer Ozdil, Yunus Emre Esin, Safak Ozturk, **Berkan Demirel**, "Representative Signature Generation for Plant Detection in Hyperspectral Images.” IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2018.
- Yunus Emre Esin, Safak Ozturk, Omer Ozdil, **Berkan Demirel**. "Registration of push broom hyperspectral camera aerial images.” Signal Processing and Communications Applications Conference (SIU), 2018 26th. IEEE, 2018.
- **Berkan Demirel**, Ramazan Gokberk Cinbis, Nazli Ikizler-Cinbis, "Attributes2Classname: A discriminative model for attribute-based unsupervised zero-shot learning.”, in IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 2017.

- **Berkan Demirel**, Yunus Emre Esin, and Ömer Özdil. "Vegetation detection with spatial segmentation and spectral indices.” Signal Processing and Communications Applications Conference (SIU), 2017 25th. IEEE, 2017.
- Omer Özdil, Yunus Emre Esin and **Berkan Demirel**. "Forming representative signature for vegetation detection in hyperspectral images.” Signal Processing and Communications Applications Conference (SIU), 2017 25th. IEEE, 2017.
- **Berkan Demirel**, Ramazan Gökberk Cinbiş, and Nazlı İkizler-Cinbiş. "Visual Saliency Estimation via Attribute Based Classifiers and Conditional Random Field.” Signal Processing and Communication Application Conference (SIU), 2016 24th. IEEE, 2016. **(Alper Atalay Best Student Paper Award)**.
- **Berkan Demirel**, Ömer Özdil, and Yunus Emre Esin. "Hyperspectral Image Segmentation Based on Spatial Model” Signal Processing and Communication Application Conference (SIU), 2016 24th. IEEE, 2016.