

DATA-EFFICIENT AND ROBUST DEEP LEARNING FOR COMPUTER VISION

PRASHANT PANDEY



**DEPARTMENT OF BHARTI SCHOOL OF
TELECOMMUNICATIONS TECHNOLOGY &
MANAGEMENT
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY, 2026**

© **Indian Institute of Technology Delhi (IITD), New Delhi, 2026**

Data-Efficient and Robust Deep Learning for Computer Vision

by

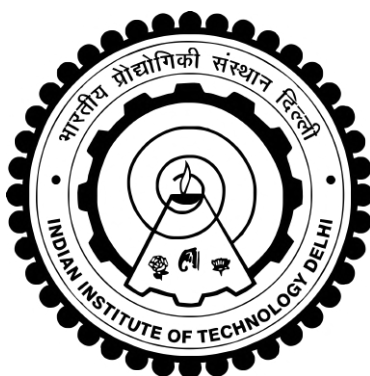
Prashant Pandey

Department of Bharti School of Telecommunications Technology &
Management

Submitted

In fulfillment of the requirements of the degree of
Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

July, 2026

CERTIFICATE

This is to certify that the thesis titled **Data-Efficient and Robust Deep Learning for Computer Vision**, submitted by **Mr. Prashant Pandey (2017BSZ8495)**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Brejesh Lall

Professor

Dept. of Electrical Engineering

IIT-Delhi, 110016

Place: New Delhi

Date:

ACKNOWLEDGEMENTS

I am very grateful to my supervisor Prof. Brejesh Lall for motivating me to do quality research and providing me with all the necessary facilities and I am very grateful to the members of the Student Research Committee and fellow collaborators for their support and constructive feedback. I am also grateful to my family members for their unwavering support and constant encouragement. I would like to thank High Performance Computing (HPC) at the Indian Institute of Technology Delhi (IITD) for providing computational resources.

Mr. Prashant Pandey

ABSTRACT

Collecting quality datasets and annotating them is a tedious process. Data collection can often be costly due to expensive devices in the field and the inaccessibility of expert annotators. During annotation, labels may contain noise, or datasets may be noisy due to inherent problems in data collection devices. In some fields like the medical domain, the annotation process is reviewed and approved at various levels, adding to the complexity of collecting quality data. Often it becomes hard to label enormous amounts of data specifically in segmentation problems where every pixel is labelled. Also, new domains or classes with distributional shifts may appear for which collecting data or labels can be extremely hard.

Approaches like transfer learning help to transfer knowledge learned from models trained on large-scale public datasets to problems with scarce datasets. Methods in domain adaptation and domain generalization handle new target domains having distributional shifts with the source domains on which the models are trained. Similarly, approaches from few-shot or zero-shot learning, learn scarcely labelled novel classes. Although these approaches tend to provide some solution to the existing problems, there is still a huge gap where existing methods or algorithms fail to deliver the required performance when the data or labels are scarce or new scarcely labelled domains or classes appear with distributional shifts or data is noisy. In this work, I have tried to solve these challenges by designing novel efficient and robust deep learning methods that can handle problems with scarce datasets or problems that require learning new scarcely labelled domains and classes with distributional shifts or noise.

Towards the end, (in the future work section), I have explored other paradigms of efficient deep learning like domain adaptive few-shot continual learning, quantization (compression) of medical diffusion models in the medical domain that will help to reduce inference time and memory requirements and time-varying multi-agent reinforcement learning. The objective of

these future works is to build efficient and robust methods that can handle catastrophic forgetting in data scarce domains with distributional shifts, compress medical diffusion models that consume significant computational resources, and enable agents to handle time-varying dynamics of the environment in multi-agent reinforcement learning by making them robust to changes.

सार

गुणवत्तापूर्ण डेटासेट एकत्र करना और उन्हें एनोटेट (लेबल) करना एक थकाऊ प्रक्रिया है। क्षेत्र में महंगे उपकरणों और विशेषज्ञ एनोटेटरों की दुर्गमता के कारण डेटा संग्रहण अक्सर महंगा हो सकता है। एनोटेशन के दौरान, लेबलों में शोर (नॉइज़) शामिल हो सकता है, या डेटा संग्रहण उपकरणों में निहित समस्याओं के कारण डेटासेट में शोर उत्पन्न हो सकता है। चिकित्सा क्षेत्र जैसे कुछ क्षेत्रों में, एनोटेशन प्रक्रिया की विभिन्न स्तरों पर समीक्षा और स्वीकृति की जाती है, जिससे गुणवत्तापूर्ण डेटा एकत्र करने की जटिलता और बढ़ जाती है।

बड़ी मात्रा में डेटा को लेबल करना अक्सर कठिन हो जाता है, विशेष रूप से सेगमेंटेशन समस्याओं में जहाँ प्रत्येक पिक्सेल को लेबल किया जाता है। इसके अलावा, वितरणात्मक बदलाव (डिस्ट्रीब्यूशनल शिफ्ट) वाले नए डोमेन या वर्ग (क्लासेज़) सामने आ सकते हैं, जिनके लिए डेटा या लेबल एकत्र करना अत्यंत कठिन हो सकता है।

ट्रांसफर लर्निंग जैसे दृष्टिकोण, बड़े पैमाने के सार्वजनिक डेटासेट पर प्रशिक्षित मॉडलों से सीखे गए ज्ञान को दुर्लभ डेटासेट वाली समस्याओं में स्थानांतरित करने में मदद करते हैं। डोमेन अडैप्टेशन और डोमेन जनरलाइज़ेशन में उपयोग की जाने वाली विधियाँ, स्रोत डोमेन की तुलना में वितरणात्मक बदलाव वाले नए लक्ष्य (टारगेट) डोमेन को संभालती हैं। इसी तरह, फ्यू शॉट या ज़ीरो शॉट लर्निंग के दृष्टिकोण, कम लेबल वाले नए वर्गों को सीखते हैं। यद्यपि ये दृष्टिकोण मौजूदा समस्याओं का कुछ हद तक समाधान प्रदान करते प्रतीत होते हैं, फिर भी एक बड़ा अंतर बना रहता है जहाँ मौजूदा विधियाँ या एल्गोरिदम आवश्यक प्रदर्शन देने में विफल रहते हैं, जब डेटा या लेबल दुर्लभ या नए हों, या वितरणात्मक बदलाव या शोर के साथ कम लेबल वाले नए डोमेन या वर्ग सामने आएँ।

इस कार्य में, मैंने ऐसे नवीन, कुशल और सुदृढ़ (रोबस्ट) डीप लर्निंग विधियों को डिज़ाइन करके इन चुनौतियों को हल करने का प्रयास किया है, जो दुर्लभ डेटासेट वाली समस्याओं को संभाल सकें, या ऐसी समस्याओं को संभाल सकें जिनमें वितरणात्मक बदलाव या शोर के साथ नए, कम लेबल वाले डोमेन और वर्गों को सीखने की आवश्यकता हो।

अंत में (भविष्य के कार्य वाले खंड में), मैंने कुशल डीप लर्निंग के अन्य प्रतिमानों (पैराडाइम्स) की खोज की है, जैसे: डोमेन अडैप्टिव फ्यू शॉट कंटीन्युअल लर्निंग, चिकित्सा क्षेत्र में मेडिकल डिफ्यूज़न मॉडलों का

क्वांटाइज़ेशन (संपीड़न) जो इन्फरेंस समय और मेमोरी आवश्यकताओं को कम करने में मदद करेगा, और मल्टी एजेंट रीइन्फोर्समेंट लर्निंग में समय परिवर्तनशील (टाइम वेरीइंग) गतिशीलता। इन भविष्य के कार्यों का उद्देश्य ऐसी कुशल और सुदृढ़ विधियाँ विकसित करना है जो वितरणात्मक बदलाव वाले डेटा दुर्लभ डोमेन में विनाशकारी विस्मरण (कैटास्ट्रॉफिक फॉरगेटिंग) को संभाल सकें, महत्वपूर्ण कम्प्यूटेशनल संसाधनों की खपत करने वाले मेडिकल डिफ्यूज़न मॉडलों को संपीड़ित कर सकें, और मल्टी एजेंट रीइन्फोर्समेंट लर्निंग में एजेंटों को वातावरण की समय परिवर्तनशील गतिशीलता के प्रति सुदृढ़ बनाकर उन्हें उससे निपटने में सक्षम बना सकें।

Contents

CERTIFICATE	i
ACKNOWLEDGEMENTS	ii
ABSTRACT	iii
LIST OF FIGURES	x
LIST OF TABLES	xii
1 INTRODUCTION	1
1.1 Transfer Learning with publicly available videos	2
1.2 Camera Invariance through Domain Adaptation for White Blood Cells . .	4
1.3 Domain Generalization with Inference-time Label-Preserving Target Projections	5
1.4 Robust Few-Shot Semantic Segmentation	6
1.5 Robust Prototypical Few-Shot Organ Segmentation with Regularized Neural-ODEs	7
1.6 Weakly Supervised Zero-Shot and Few-Shot Semantic Segmentation with Mean Instance Aware Prompt Learning	9
2 RELATED WORK	11
2.1 Transfer Learning with publicly available videos	11
2.1.1 Use of Technology in Autism	11
2.1.2 Human Action Recognition from Videos	12
2.2 Camera Invariance through Domain Adaptation for White Blood Cells . .	13
2.2.1 Use of Technology in White Blood Cells (WBC) Classification . . .	13
2.2.2 Unsupervised Domain Adaptation for Classification	13
2.3 Domain Generalization for Classification	16
2.4 Robust Few-Shot Semantic Segmentation	18

2.4.1	Neural-ODEs	18
2.4.2	Few-Shot Semantic Segmentation	19
2.4.3	Adversarial robustness	20
2.5	Weakly Supervised Zero-Shot and Few-Shot Semantic Segmentation with Mean Instance Aware Prompt Learning	22
2.5.1	Zero-Shot Semantic Segmentation	22
2.5.2	Weakly Supervised Semantic Segmentation (WSS)	22
2.5.3	Semantic Embeddings and Language Models	23
2.5.4	Zero/Few-Shot Semantic Segmentation with Weak Supervision	23
3	ASSESSMENT OF CHILDREN WITH AUTISM SPECTRUM DISORDER: A TRANSFER LEARNING APPROACH	25
3.1	Introduction	25
3.2	Transfer Learning with publicly available videos	26
3.2.1	Guided Weak Supervision	27
3.2.2	Directional Regularization	29
3.2.3	Implementation	31
3.3	Autism Dataset	33
3.4	Experiments and Results	36
3.4.1	Baselines	36
3.4.2	Bias in the Training Data	39
3.4.3	GWS and DR under Different Settings	39
3.4.4	Comparisons with State-of-the-art	40
3.5	Ethics and Social Impact	41
3.6	Summary	42
4	DOMAIN ADAPTATION FOR CLASSIFICATION OF WHITE BLOOD CELLS	43
4.1	Introduction and Motivation	43
4.2	Existence of a Closest Source Clone	45
4.3	Variational Inference for Source Sampling	46
4.3.1	Variational Auto-Encoders	46
4.3.2	Proposed Enhancements to the VAE	47

4.4	Finding the Closest-Clone Through Latent Search	50
4.4.1	Problem Formulation	50
4.4.2	Choice of Distance Metric: SSIM Loss	50
4.4.3	The Inference Procedure	51
4.5	Dataset Details	51
4.5.1	Cell Types and Classification Taxonomy	51
4.5.2	Imaging Hardware and Data Acquisition	53
4.6	Implementation Details	54
4.6.1	VAE Training Configuration	54
4.6.2	Latent Search Configuration	54
4.7	Experiments and Results	56
4.7.1	Comparison with State-of-the-Art Methods	56
4.7.2	Ablation Studies	58
4.8	Discussion	60
4.9	Summary	61
5	DOMAIN GENERALIZATION WITH INFERENCE-TIME LABEL-PRESERVING TARGET PROJECTIONS	63
5.1	Introduction	63
5.2	Label-Preserving Transformation	64
5.2.1	Domain invariant Features	64
5.2.2	Learning the f -function	67
5.3	Inference-time Target Projections	68
5.3.1	Generating the Source-Feature Manifold $\mathcal{Z}_s$	68
5.3.2	Label-preserving Projections	70
5.4	Results	71
5.4.1	Multi-source Domain Generalization	71
5.4.2	Ablation Studies	73
5.4.3	Low resource settings	74
5.5	Summary	76
6	ROBUST FEW-SHOT SEMANTIC SEGMENTATION	77

6.1	Introduction	77
6.2	Problem setting	78
6.3	Adversarial Training	79
6.4	Prototypical Neural ODE (PNODE)	81
6.5	Regularized Prototypical Neural-ODE (R-PNODE)	84
6.6	Results	86
6.6.1	Prototypical Neural-ODE (PNODE)	86
6.6.2	Regularized Prototypical Neural-ODE (R-PNODE)	88
6.7	Summary	96
7	WEAKLY SUPERVISED ZERO-SHOT AND FEW-SHOT SEMANTIC SEGMENTATION WITH MEAN INSTANCE AWARE PROMPT LEARNING	97
7.1	Introduction	97
7.2	Pseudo-label Generation Module	98
7.3	Class-Agnostic Mask Generation (CAMG)	100
7.4	CLIP language model	100
7.5	Mask Aggregation	102
7.6	Generalization bound	103
7.7	Datasets	104
7.8	Implementation Details	105
7.8.1	Pascal VOC & COCO	105
7.8.2	BCSS-WSSS	106
7.9	Results	106
7.9.1	Ablation Studies	111
7.10	Summary	112
8	CONCLUSION AND FUTURE WORK	113
8.1	Future Work	113
8.1.1	Generalized Continual Learning - 2D data and beyond	114
8.1.2	Quantization of Medical Diffusion models	117
8.1.3	Time-varying Multi-agent Reinforcement Learning	117

Bibliography	121
Publications	145
Biography	147

List of Figures

3.1	Illustration of the similarity between the action classes in two different datasets in the optical flow domain.	27
3.2	Guided weak supervision (GWS) and Directional Regularization (DR). . .	28
3.3	Performance on I3D and TSN with different baselines and different amounts of augmented source (Kinetics and HMDB51) samples using GWS and DR.	34
3.4	Training sessions of Asian and Caucasian subjects by clinicians with different camera positions.	35
3.5	t-SNE plots of embeddings of penultimate layers of baseline I3D and TSN Autism models with and without Guided Weak Supervision (GWS) and Directional Regularization (DR).	36
3.6	Performance on I3D and TSN with iteration over the source samples with Guided Weak Supervision (GWS) and Directional Regularization (DR). . .	38
4.1	Training the VAE on the source WBC dataset.	49
4.2	Latent Search inference procedure for WBC classification.	52
4.3	Representative WBC and related microscopic images across three cameras.	53
4.4	Progressive transformation during Latent Search in TIGDA.	58
4.5	LS performance versus iterations for three objective functions.	59
4.6	Component-wise ablation images for the C→A task.	60
5.1	Overall procedure of the proposed method for Domain Generalization. . . .	69
5.2	(left to right) (a) Ablation on Office-Home dataset highlighting the relative importance of different components employed in our method. (b) Relative performance of different methods trained on 20%, 40%, 60% and 80% of PACS dataset. For each setting, we report average performance with all examples from the target domains using leave-one-out strategy. (c) Relative performance of DG methods trained on a single domain (Sketch (S)) and tested on Photo (P), Cartoon (C) and Art (A) domains.	74
6.1	Robust features for query and support images are obtained by the feature extractor followed by the continuous dynamics and integral solutions of a Neural-ODE.	80

6.2	Visualisation of the evolution of the features learnt across the layers in a CNN and across time for a Neural-ODE.	82
6.3	Robust-Prototypical Neural-ODE.	83
6.4	Performance of models for attacks on BCV $\rightarrow$ CT-ORG (left), BCV $\rightarrow$ Decathlon (middle) and for different intensities of FGSM (right), on Liver 1-shot.	87
6.5	Predicted masks by different models for different attacks.	88
6.6	3-shot FSS results for BCV $\rightarrow$ BCV (in-domain), BCV $\rightarrow$ CT-ORG and BCV $\rightarrow$ Decathlon for Liver organ (novel class).	90
6.7	t-SNE plots for query attack on a baseline method and R-PNODE.	91
6.8	3-shot FSS results for BCV $\rightarrow$ BCV (in-domain) and BCV $\rightarrow$ Decathlon for Liver organ (novel class) when support and query are attacked.	93
7.1	An overview of weakly supervised few-shot and zero-shot semantic segmentation.	97
7.2	Overview of the proposed mean instance aware prompt learning.	99
7.3	Comparison of the proposed prompt learning strategy with prior works.	101
7.4	Predicted masks with different prompt learning strategies in weak generalized zero-shot segmentation (WGZSS) setting on PASCAL VOC.	108
7.5	1-way 1-shot predicted masks on Pascal-5 ⁱ	109
7.6	1-way 1-shot predicted masks on COCO-20 ⁱ	109
7.7	2-way 1-shot predicted masks on COCO-20 ⁱ	110
7.8	Semantic segmentation on Stroma class.	110
7.9	Ablation with different prompting strategies.	111
7.10	t-SNE plots showing image and text features obtained from different prompt learning methods. The left figure is obtained using Xu et al. [2022a] and the right from MIAPNet.	111
8.1	Low-bit (2/4 bit) post-training quantization of a diffusion model on the Endora dataset that generates endoscopy videos.	117
8.2	Effect of variation of gravity on agents in a Mujoco robotic environment Cheetah.	118
8.3	Effect of variation of reward on agents in a Mujoco robotic environment Walker.	119
8.4	Effect of variation of noise in the observation space on agents in the SMAC environment.	120

List of Tables

3.1	Action classes from the source datasets (Kinetics and HMDB51) matched to target (Autism) classes.	37
3.2	Performance of GWS and DR with a specific bias in the Autism training dataset.	38
3.3	GWS and DR under different settings.	38
3.4	Comparison with state-of-the-art action recognition models on Autism dataset and SSBD dataset.	39
4.1	Class-wise image counts for each of the three cameras.	54
4.2	Accuracy (mean $\pm$ std%) for UDA tasks on WBC images from three cameras.	57
4.3	Ablation of TIGDA components on two cross-camera tasks.	60
5.1	Comparison of performance between different models on PACS Li et al. [2017] dataset with AlexNet, ResNet-18 and ResNet-50 as backbones for the f_θ network.	72
5.2	Comparison of performance between different models using AlexNet backbone on VLCS Fang et al. [2013] dataset.	72
5.3	Comparison of performance between different models using ResNet-18 backbone on Office-Home Venkateswara et al. [2017].	72
5.4	Comparison of performance between different models on Digits-DG Zhou et al. [2020b] dataset.	73
5.5	Comparison of few-shot domain adaptation performance between different models on Digits dataset (MNIST LeCun et al. [1998] (M), USPS Hull [1994] (U) and SVHN Netzer et al. [2011] (S)).	75
5.6	Ablation on Single Source DG with Sketch (S) as source and Photo (P), Cartoon (C) and Art (A) as unseen target domains.	76
6.1	1-shot query attack results for BCV $\rightarrow$ BCV in-domain Liver and Spleen organs (novel classes).	86
6.2	3-shot query attack results for Liver BCV in-domain and BCV $\rightarrow$ CT-ORG cross-domain settings.	87
6.3	Few-shot organ segmentation results on 1-shot setting.	89

6.4	Query attack results (higher is better) for BCV $\rightarrow$ BCV 1-shot in-domain for Liver organ (novel class).	91
6.5	Query attack results (higher is better) for BCV $\rightarrow$ BCV 1-shot in-domain for Spleen organ (novel class).	91
6.6	Query attack results for BCV $\rightarrow$ CT-ORG for Liver. The dice scores are rounded off to two decimals.	92
6.7	Query attack results for BCV $\rightarrow$ Decathlon for Liver. Dice scores are rounded off to two decimals.	92
6.8	Support attack results with prototypical networks for BCV $\rightarrow$ BCV in-domain for Liver and Spleen as novel classes.	94
6.9	Support attack results with prototypical networks for BCV $\rightarrow$ CT-ORG and BCV $\rightarrow$ Decathlon settings for Liver.	94
6.10	Ablation studies of different components of R-PNODE for BCV $\rightarrow$ BCV in-domain for Spleen organ.	95
7.1	Generalized zero-shot semantic segmentation (GZSS) results on PASCAL VOC 2012.	107
7.2	Zero-shot semantic segmentation (ZSS) results on Pascal-5 ⁱ	107
7.3	ZSS results on COCO-20 ⁱ	108
7.4	1-way 1-shot few-shot semantic segmentation (FSS) results on Pascal-5 ⁱ	108
7.5	1-way 1-shot FSS results on COCO-20 ⁱ	109
7.6	Semantic segmentation on Stroma class.	110
8.1	Few-shot continual learning for semantic segmentation.	114
8.2	Few-shot continual speaker adaptation for co-speech gesture generation.	116