Enhancing Generalization in Sketch-Based Image Retrieval through Single and Multi-Source Domain Adaptation

by

Mengqing Huang

National Centre for Computer Animation

Faculty of Media & Communication

Bournemouth University

A thesis submitted in partial fulfilment of the
requirements of Bournemouth University for the degree of
Doctor of Philosophy

Jan. 2024

Copyright Statement

Acknowledgements

I wish to express my sincere thanks to my supervisors, Dr Hongchuan Yu and Professor Jian J. Zhang, for their continuous support of my PhD study and research. Their insight and expertise led me through the nebulous path to becoming a researcher. I would particularly like to single out my supervisor Professor Hongchuan Yu. He participated in every crunch, guiding me to embrace the wildest vision for research while breaking down problems to the finest details. Without his guidance and encouragement this thesis would never have materialised.

I would like to thank Dr. Jinglu Zhang, Dr. Yunfei Fu, Dr. Yanran Li, Jiajun Huang who offered unconditional help on various aspects in my research. I also would like to thank Sonia Ashby, Cansu Kurt Green and George Askwith who handled all the administrative tasks and cared my student life in Bournemouth University.

In addition, I would like to thank my awesome friends Dr. Yu Xia, Dr. Yao Lyu, Dr. Shuangbu Wang, Dr. Nan Xiang, Bing Yang, Dr. Yinyu Nie, Dr. Zhangcan Ding, Dr. Ruibin Wang, Yu Zhou, Dr. Tao Jiang, Junheng Fang, Dr. Alexander Tereshin, Dr. Mara Catalina Aguilera Canon, Kavisha Jayathunge, Xingyu Ye, Yingjie Xi for the invaluable friendship.

I thank my parents for their continued love and support. I am forever grateful for my caring, patient, and supportive family. I would also like to thank my girlfriend Xiaoxiao Liu who has been my steadfast support and source of encouragement throughout this journey.

Abstract

This thesis addresses the critical challenge of generalization in deep neural networks, particularly within the context of Sketch-Based Image Retrieval (SBIR). A primary contribution is the development of a novel empirical framework to evaluate and benchmark the generalization capacity of deep networks. This framework introduces metrics that quantify both model accuracy and the ability to handle data diversity, offering a practical approach to assess model performance on unseen data and identifying trade-offs crucial for effective model selection.

Building on this generalization framework, the research proposes domain adaptation strategies specifically tailored for SBIR to bridge the significant gap between sketch and image domains . A single-source domain adaptation algorithm is introduced, utilizing canonical correlation analysis (CCA) alongside dictionary learning principles and sparse optimization techniques to facilitate effective knowledge transfer from a source (e.g., images) to a target domain (e.g., sketches), even in few-shot scenarios . This approach is further extended to a multi-source domain adaptation algorithm, capable of integrating information from multiple diverse source domains to enhance robustness and adaptability. Computational efficiency is a key consideration, addressed through the use of low-rank matrix decomposition and online dictionary learning techniques.

Overall, this work provides a comprehensive approach to enhancing SBIR system performance by directly tackling generalization limitations through principled empirical assessment and efficient

single- and multi-source domain adaptation methods. The findings contribute to both the understanding of generalization in deep learning and the development of practical, adaptable SBIR systems.

Contents

List of Figures

List of Tables

List of Abbreviations

SBIR Sketch-based Image Retrieval

ZS-SBIR Zero-Shot Sketch-based Image Retrieval

DA-SBIR Domain Adaptation Sketch-Based Image Retrieval

PGDL Predicting Generalization in Deep Learning

GAN Generative Adversarial Network

DGMs Deep Generative Models

ZSL Zero-Shot Learning

CNN Convolutional Neural Network

CBIR Content-based Image Retrieval

Chapter 1

Introduction

In the realm of deep learning, the quest for understanding and optimizing model performance transcends theoretical boundaries and delves into diverse practical applications. This thesis focuses on enhancing the generalization capabilities of Sketch-Based Image Retrieval (SBIR) systems through effective domain adaptation strategies. By developing methods to quantitatively assess generalization and implementing domain adaptation techniques specifically tailored for SBIR, this research aims to advance the state of the art in sketch-based retrieval systems.

1.1 Background and Motivation

Sketch-based Image Retrieval (SBIR) stands at the intersection of computer vision, multimedia analysis, and human-computer interaction, offering an intuitive approach to search for images using hand-drawn sketches. The underlying premise of SBIR is compelling: humans can communicate visual concepts through simple sketches that capture essential structural attributes, allowing for a more natural query mechanism than text-based alternatives. The capacity to search using free-hand sketches offers significant advantages in scenarios where textual descriptions prove inadequate, or when the searcher's mental image is more readily articulated visually than verbally.

The origins of SBIR can be traced back to the early 1990s with pioneering works by (Faloutsos et al. 1994), (Hirata and Kato 1992), and (Niblack et al.

1993). These early systems primarily focused on simple contour matching between line drawings and image edges. However, the field remained relatively constrained due to limited computational resources and the scarcity of digital sketch data. The landscape transformed dramatically with the widespread adoption of touchscreen devices, enabling the collection of large sketch datasets and facilitating more sophisticated SBIR approaches.

The remarkable growth in SBIR research coincided with broader advances in computer vision and machine learning, particularly deep learning. Modern SBIR systems leverage convolutional neural networks (CNNs) and other deep architectures to learn discriminative representations that bridge the substantial gap between sketches and photographic images. This evolution has shifted the focus from handcrafted feature engineering to data-driven representation learning, significantly enhancing retrieval performance.

Despite these advancements, SBIR systems continue to face significant challenges that limit their practical application. A fundamental issue lies in the inherent domain gap between sketches and photos – while sketches primarily capture shape and structural information through sparse line drawings, photographs contain rich texture, color, and contextual details. Furthermore, sketches exhibit high variability due to differences in drawing style, abstraction level, and artistic ability across users. These factors create a complex retrieval problem that demands robust and adaptive solutions.

Additionally, the rapid development of data generation and the diversity of visual content necessitate SBIR systems that can generalize beyond their training distributions. This requirement becomes particularly challenging when confronted with new categories or domains not represented in the training data, a scenario known as zero-shot or few-shot learning. The ability to adapt to novel classes or domains without extensive retraining represents a critical frontier in SBIR research and forms a core motivation for the work presented in this thesis.

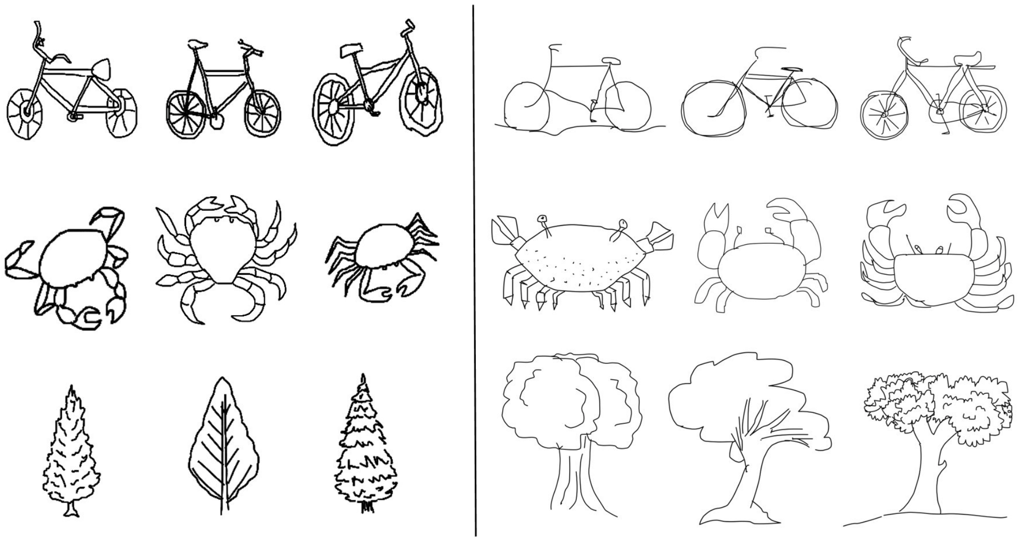


Figure 1.1: The representative sketches of (a) the Sketchy Dataset and (b) the TU-Berlin dataset.

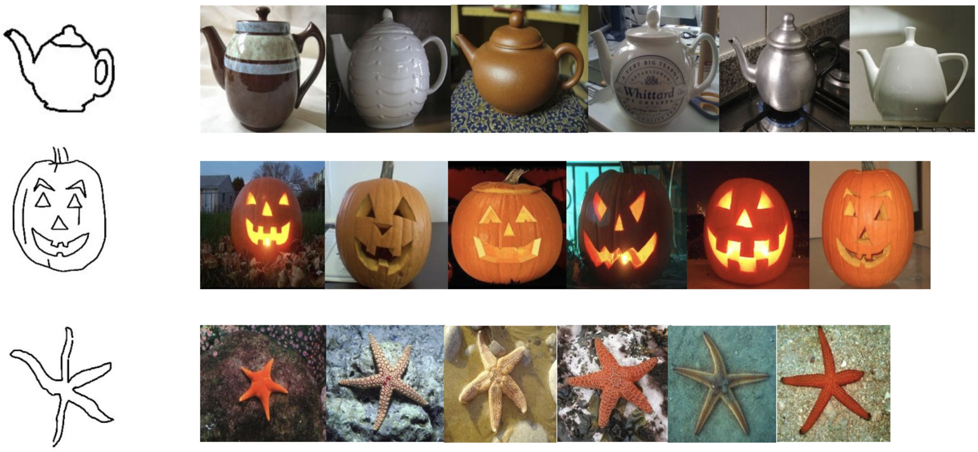


Figure 1.2: The sketch-based image retrieval examples

1.2 Research Problems

The fundamental research problem addressed in this thesis centers on enhancing the generalization capabilities of Sketch-Based Image Retrieval systems through domain adaptation. While current SBIR approaches demonstrate reasonable performance within controlled settings using well-established datasets, they often falter when confronted with data from novel domains or categories. This limitation significantly hampers their practical utility in real-world applications where the diversity and continual evolution of visual content are inevitable.

The research problem encompasses several interconnected challenges:

1. **Domain Gap**: A substantial gap exists between the domain of sketches and that of photographic images, characterized by differences in visual characteristics, information density, and abstraction levels. Existing SBIR systems struggle to bridge this gap effectively, particularly for categories with high intra-class variability.

2. **Limited Generalization**: Current SBIR models demonstrate inadequate generalization to unseen data, manifesting as performance degradation when tested on categories or visual styles not represented in the training data. This limitation restricts the practical applicability of these systems in dynamic, real-world scenarios.

3. **Dataset Discrepancies**: Multiple SBIR datasets exist with overlapping categories but distinct characteristics due to variations in collection methodologies, sketching styles, and abstraction levels. The differences between these datasets introduce additional complexity when attempting to leverage diverse data sources.

4. **Few-Shot Learning Constraints**: In practical scenarios, new visual categories emerge continuously, but obtaining substantial labeled data for these categories is resource-intensive. SBIR systems must therefore adapt to new categories with minimal labeled examples—a capability not adequately addressed by existing approaches.

5. **Computational Efficiency**: The computational demands of sophisticated deep learning models pose practical challenges for deployment, particularly in resource-constrained environments. Efficient domain adaptation techniques that maintain performance while reducing computational overhead are essential.

1.3 Research Questions

This thesis addresses the following key research questions:

1. **How can the generalization capacity of deep neural networks for SBIR be effectively quantified?** This question examines the development of metrics that accurately reflect a model's ability to perform well on unseen data, particularly across different domains. It explores the relationship between model accuracy, data diversity, and generalization performance.

2. **What methods can effectively bridge the domain gap between sketches and photographic images while maintaining computational efficiency?** This question investigates adaptation techniques that preserve domain-invariant information while accommodating the unique characteristics of sketches, such as abstraction and structural emphasis.

3. **How can knowledge from multiple source domains be integrated to enhance SBIR performance in few-shot and zero-shot scenarios?** This question explores the leveraging of diverse visual domains (e.g., clipart, real images, paintings) to improve retrieval performance, particularly when limited labeled data is available for target categories.

4. **To what extent do traditional machine learning techniques complement deep learning approaches in addressing domain adaptation for SBIR?** This question examines the synergistic potential of integrating established techniques such as Canonical Correlation

Analysis with deep learning models to overcome the limitations of pure deep learning approaches.

5. **What is the relationship between model complexity, computational efficiency, and retrieval performance in domain-adapted SBIR systems?** This question explores the trade-offs between model sophistication, computational requirements, and retrieval accuracy, seeking optimal configurations for practical deployment.

These research questions guide the investigations presented in this thesis and provide a framework for evaluating the contributions of the developed methods.

This thesis contends that addressing these challenges requires a comprehensive approach that combines robust generalization metrics with effective domain adaptation strategies. By developing methods to quantitatively assess generalization capabilities and implementing domain adaptation techniques specifically tailored for SBIR, this research aims to advance the state of the art in sketch-based retrieval systems.

1.4 Research Aims and Objectives

The overarching aim of this research is to develop a robust framework for enhancing the generalization capabilities of Sketch-Based Image Retrieval systems through effective domain adaptation strategies. This aim is pursued through the following specific objectives:

1. **Develop an empirical generalization metric for deep networks**: Establish a quantitative framework to assess the generalization capabilities of deep neural networks used in SBIR, enabling objective comparison between different architectures and identifying optimal models for cross-domain retrieval tasks.

2. **Design and implement single-source domain adaptation methods for SBIR**: Create adaptation techniques that effectively transfer

knowledge from a source domain (photographic images) to a target domain (sketches), bridging the fundamental domain gap in SBIR while maintaining computational efficiency.

3. **Extend domain adaptation capabilities to multi-source scenarios**: Develop methods that leverage multiple source domains simultaneously to enhance the robustness and versatility of SBIR systems, particularly for handling diverse visual representations beyond the conventional sketch-photo pair.

4. **Evaluate and validate the proposed methods on benchmark datasets**: Conduct comprehensive experimentation on established SBIR datasets to quantify improvements in retrieval performance, particularly focusing on few-shot and zero-shot scenarios that reflect real-world challenges.

5. **Analyze the computational efficiency of the proposed approaches**: Assess the computational requirements of the developed methods to ensure practical applicability, particularly through the use of low-rank matrix decomposition and other efficiency-enhancing techniques.

These objectives collectively address the identified research problems and contribute to advancing the field of SBIR, with particular emphasis on enhancing generalization capabilities through domain adaptation.

1.5 Scope and Limitations

This research focuses specifically on domain adaptation for Sketch-Based Image Retrieval using single and multiple source domains. While comprehensive within this scope, several important limitations delineate the boundaries of this work:

1. **Focus on 2D Sketches and Images**: This research exclusively addresses 2D sketches and photographic images, excluding 3D models, videos, or other multimedia formats that might benefit from sketch-based retrieval.

2. **Categorical Retrieval**: The retrieval paradigm employed is primarily categorical, aiming to retrieve images from the same semantic category as the query sketch, rather than focusing on fine-grained instance-level retrieval.

3. **Pre-drawn Sketches**: The evaluation utilizes pre-drawn sketches from established datasets rather than real-time sketches drawn by users in interactive settings, which might introduce additional variations and timing considerations.

4. **Computational Constraints**: The methods developed prioritize practical computational efficiency, potentially sacrificing theoretical optimality for approaches that can be feasibly implemented in realistic settings.

5. **Dataset Limitations**: The evaluation is conducted on established benchmark datasets (Sketchy, TU-Berlin, DomainNet) which, while comprehensive, may not entirely reflect the diversity of real-world sketching styles and image types.

These limitations establish a focused research scope while acknowledging potential areas for future expansion beyond the current work.

1.6 Thesis contributions and outlines

In this thesis, I introduce a pioneering approach to addressing the critical aspect of generalization in deep learning, which forms the foundation for the subsequent exploration in sketch-based image retrieval (SBIR). The main contributions are summarized as follows:

Framework for Assessing Generalization in Deep Learning: The foremost contribution is the development of an innovative framework that introduces an intuitive metric system for benchmarking deep networks. This system, focusing on both model accuracy and the diversity of unseen data, provides a comprehensive assessment of a network's generalization capacity. It offers vital quantitative and qualitative insights, paving the way for understanding and enhancing the generalization capabilities of deep neural networks across various learning scenarios.

Building upon this understanding of generalization, the research extends to specifically address the Domain Adaptation (DA) problem in practical SBIR scenarios:

1. **Single Source DA-SBIR Algorithm**: Leveraging insights from the generalization framework, I propose an algorithm for image-to-sketch domain adaptation and dataset-to-dataset domain adaptation. This algorithm facilitates the transfer of learning models from one source domain to a target, enhancing adaptability based on generalization principles.

2. **Multiple Sources DA-SBIR Algorithm**: Expanding further, I introduce an algorithm that transfers learning models to a target domain using insights gained from multiple source domains. This approach underscores the importance of generalization across varied domains.

3. **Efficiency in Computational Complexity**: In line with the generalization-centric approach, I employ canonical correlation analysis and online dictionary learning technologies. These technologies are chosen for their ability to handle large datasets efficiently, aligning with the need for effective generalization in complex scenarios.

By emphasizing the importance of generalization in deep learning, this thesis lays a robust foundation for addressing the DA problem in SBIR. The proposed DA-SBIR algorithms utilize low-rank matrix decomposition technology for its efficiency, and the methodologies extend to zero-shot settings, applying deep learning techniques informed by the generalization framework. This comprehensive approach underlines the interconnectivity between generalization in deep learning and the specific challenges in SBIR.

1.7 Thesis Structure

This thesis is organized into six chapters, each addressing specific aspects of the research on enhancing generalization in Sketch-Based Image Retrieval (SBIR) through domain adaptation: As illustrated in Figure 1.3, the thesis follows a logical progression that maps to the components of the SBIR system:

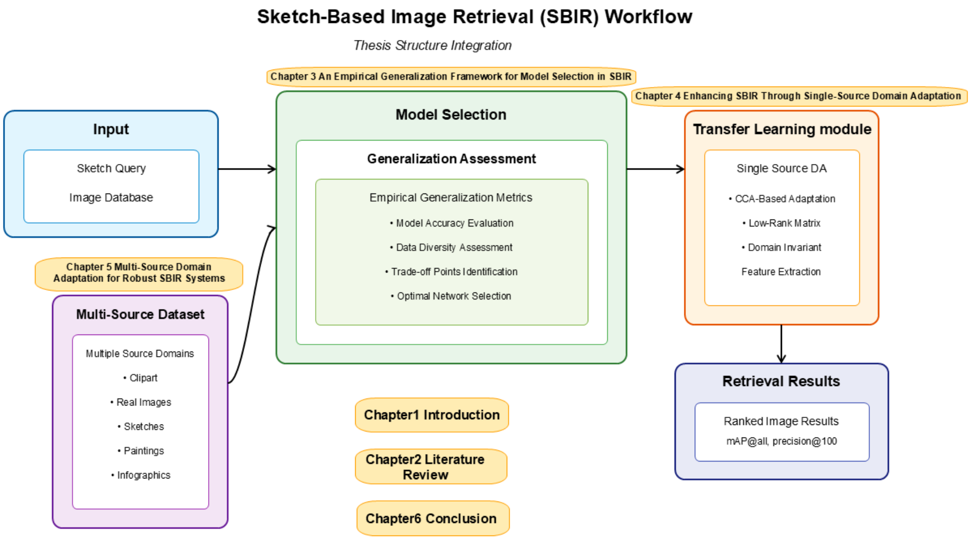

Sketch-Based Image Retrieval (SBIR) Workflow
Thesis Structure Integration
Chapter 3 An Empirical Generalization Framework for Model Selection in SBIR
Chapter 4 Enhancing SBIR Through Single-Source Domain Adaptation
Input
Sketch Query
Image Database
Model Selection
Generalization Assessment
Empirical Generalization Metrics
• Model Accuracy Evaluation
• Data Diversity Assessment
• Trade-off Points Identification
• Optimal Network Selection
Transfer Learning module
Single Source DA
• CCA-Based Adaptation
• Low-Rank Matrix
• Domain Invariant
Feature Extraction
Chapter 5 Multi-Source Domain Adaptation for Robust SBIR Systems
Multi-Source Dataset
Multiple Source Domains
• Clipart
• Real Images
• Sketches
• Paintings
• Infographics
Chapter1 Introduction
Chapter2 Literature Review
Chapter6 Conclusion
Retrieval Results
Ranked Image Results
mAP@all, precision@100

Figure 1.3: The SBIR workflow showing the relationship between thesis chapters and system components. Chapter 3 develops the foundation model selection methodology, Chapter 4 implements single-source domain adaptation, and Chapter 5 extends this to multi-source scenarios.

- Chapter 2 reviews the literature on generalization in deep learning and state-of-the-art approaches to sketch-based image retrieval. It provides theoretical foundations and identifies research gaps in domain adaptation for SBIR that this thesis addresses.

- Chapter 3 introduces a framework for assessing generalization in deep learning, establishing a methodology for selecting optimal foundation models. This chapter develops empirical metrics that quantify both model accuracy and data diversity, providing the theoretical basis for the model selection component shown in the workflow.

- Chapter 4 addresses the transfer learning component through single-source domain adaptation for SBIR. It presents methods for effectively transferring knowledge from a source domain (photographic images) to a target domain (sketches), implementing the domain adaptation techniques shown in the central flow of the workflow diagram.
- Chapter 5 extends the domain adaptation capabilities to multi-source scenarios, enabling the system to leverage diverse input domains simultaneously. This chapter corresponds to the input expansion component in the workflow, allowing the system to handle multiple visual domains beyond the conventional sketch-photo pair.
- Chapter 6 concludes the thesis, summarizing the main findings, discussing limitations, and suggesting directions for future research to further enhance generalization in SBIR systems.

The integration of these components creates a comprehensive SBIR system with enhanced generalization capabilities through effective domain adaptation, implementing the complete workflow from foundation model selection through single-source and multi-source domain adaptation to produce improved retrieval results.

1.8 Previously published work

Chapter 3 contains work published in Scientific Reports, 2025. This publication presents the deep network benchmarking and encourage contributions to expand the dataset and foster further theoretical and practical research.

Chapter 4 and Chapter 5 contain work published in ACM Transactions on Multimedia Computing, Communications and Applications, 2023. This publication presents the domain adaptation methods for both single-source and multi-source scenarios in SBIR, forming a substantial component of the contributions presented in this thesis.

Chapter 2

Literature Review

This chapter presents a comprehensive review of literature related to generalization metrics, sketch-based image retrieval, and domain adaptation techniques. Following the workflow established in Chapter 1, this review is structured to provide the theoretical foundation for the methodologies developed in subsequent chapters. The review begins with an examination of generalization in deep learning networks, which forms the foundation for model selection. It then explores sketch-based image retrieval approaches and the challenges they face, before delving into domain adaptation methods that bridge the gap between sketches and images. This progression establishes a coherent narrative that connects the components of the SBIR workflow: from foundation model selection to single and multi-source domain adaptation.

2.1 Generalization in Deep Learning

Generalization—the ability of a model to perform effectively on unseen data—is a cornerstone concept in machine learning. This section explores recent advances in understanding and measuring generalization in deep neural networks.

2.1.1 Theoretical Frameworks for Generalization Bounds

Traditional machine learning theories, primarily based on worst-case scenarios as noted by Zhang et al. (2021b), have proven insufficient in comprehensively explaining the generalization observed in deep learning models. This gap is particularly pronounced in understanding the robust generalization performance of over-parameterized neural networks, as discussed by Neyshabur et al. (2018).

A pivotal contribution to this field was made by Neyshabur et al. (2018), who introduced a complexity measure grounded in unit-wise capacities, providing a refined generalization bound for two-layer ReLU networks. Valle-Pérez and Louis (2020) further expanded on this topic through a detailed review of generalization error bound estimation. Their review proposed seven criteria for evaluating generalization in deep learning models, systematically categorizing existing approaches based on these criteria.

These approaches can be categorized into four main groups:

1. **Data-independent and algorithm-independent algorithms**: Characterized by minimal assumptions and low reliance on training data. This approach, exemplified by the VC dimension bounds explored in studies like those by Shalev-Shwartz and Ben-David (2014) and Harvey et al. (2017), represents a foundational approach in the field.

2. **Data-dependent but algorithm-independent algorithms**: These methods rely on training data while maintaining minimal assumptions about the models. Key examples include the Rademacher complexity bounds, as discussed in the works of Bartlett and Mendelson (2002) and Shawe-Taylor and Williamson (1997).

3. **Data-independent but algorithm-dependent algorithms**: These approaches, including those by Hardt et al. (2016), Mou et al. (2018), and Brutzkus et al. (2017), carry strong assumptions about the models but do not depend on the specifics of the training data.

4. **Data-dependent and algorithm-dependent algorithms**: Characterized by strong assumptions that rely heavily on the training data. This approach is evident in the methodologies proposed by Barron and Klusowski (2019), Golowich et al. (2018), and Neyshabur et al. (2018), among others.

These diverse research efforts are crucial not only for their theoretical significance but also for laying the groundwork for a deeper understanding of generalization in deep learning.

2.1.2 Generalization in Generative Models

Beyond supervised learning, Generative Adversarial Networks (GANs) have gained prominence in modeling complex real-world data. A noteworthy observation by Radford et al. (2021) suggests that GANs tend to produce synthetic datasets that align more closely with test sets than training sets in well-trained deep network classifiers' feature spaces. This finding points to the potential of GANs in exploring generalization error bounds, although evaluating the generalization capacity of Deep Generative Models (DGMs) remains challenging due to dimensionality issues.

Metrics for evaluating DGMs, such as the Inception Score and Fréchet Inception Distance, aim to estimate the distance between generated and target distributions using a polynomial number of samples. Despite their intuitive appeal and computational efficiency, the reliability of these metrics has been called into question. Addressing these concerns, Thanh-Tung and Tran (2020) introduced a Minimum Description Length-inspired metric suitable for a broad class of generative latent variable models.

The NeurIPS 2020 Predicting Generalization in Deep Learning competition (Jiang et al. 2020) provided valuable insights into the correlation between model complexity and actual generalization errors using Conditional Mutual Information. This competition highlighted the difficulty in developing reliable metrics for predicting generalization performance, emphasizing the need for empirical approaches that capture a broad range of hyperparameter variations.

2.2 Sketch-Based Image Retrieval

Sketch-based Image Retrieval (SBIR) has evolved significantly over the past decade, transitioning from traditional feature-based methods to sophisticated deep learning approaches. This section examines this evolution and the current state of the art in SBIR.

2.2.1 Traditional SBIR Methods

Early SBIR approaches relied on hand-crafted features and traditional machine learning techniques. These methods typically operated within a bag-of-words search framework (Hu and Collomosse 2013, Saavedra 2014), focusing on matching edge maps extracted from photos with input sketches. While conceptually straightforward, these approaches faced significant challenges in handling the inherent abstraction and variability in sketches, as well as the substantial domain gap between sketches and photographs.

Traditional SBIR methods generally followed a pipeline of:

1. Edge detection and feature extraction from images
2. Shape/contour matching between sketch features and image edge maps
3. Similarity computation and ranking

Despite their intuitive design, these methods struggled with real-world variations in sketching styles and the semantic gap between abstract sketches and detailed photographs.

2.2.2 Deep Learning Approaches for SBIR

The advent of deep learning has revolutionized SBIR, enabling more robust feature representations and better cross-domain matching. Contemporary SBIR methods can be broadly categorized into two approaches:

1. **Representation Learning**: Methods that learn discriminative representations for sketches and images, often using convolutional neural networks (CNNs). SketchNet (Zhang et al. 2016) pioneered this direction by employing a triplet network architecture consisting of sketch, positive, and negative photos for training their deep model, showing significant improvements in retrieval with deep feature representation.

2. **Cross-Modal Matching**: Approaches that explicitly model the relationship between sketches and images, often using siamese or triplet networks with specialized loss functions. Works like Yu et al. (2016), Song et al. (2017), and Li et al. (2017) focus on identifying subtle differences between photos and sketches, addressing challenges such as the high abstraction of sketches and the scarcity of labeled sketch-photo datasets.

Deep hashing methods have also gained prominence in SBIR, enabling efficient retrieval at scale. Efforts like Deep Sketch Hashing (Liu et al. 2017) and Generative Domain-migration Hashing (Zhang et al. 2018) have been developed to align sketches with visually similar images while preserving sketch-specific details.

Recent research has further refined these approaches by incorporating additional modalities. For instance, Dutta and Akata (2020) and Wang et al. (2021) have integrated textual semantic information into deep networks to address the domain adaptation challenges in SBIR.

2.3 Zero-Shot and Few-Shot Learning in SBIR

The practical utility of SBIR systems hinges on their ability to handle new categories with minimal or no labeled examples, a scenario commonly addressed through zero-shot and few-shot learning frameworks. This section explores these approaches in the context of SBIR.

2.3.1 Zero-Shot Learning for SBIR

Zero-shot learning (ZSL) in SBIR refers to the capability of retrieving images from categories that were not seen during training. This capability is crucial for practical applications, as it is impractical to train models on all possible object categories. A comprehensive review of ZSL approaches is provided by Xian et al. (2018).

ZSL methods in SBIR generally rely on indirect associations, particularly through semantic spaces, to bridge seen and unseen categories. The progression of ZSL approaches in the computer vision field provides valuable insights for SBIR:

1. **Early Attribute-Based Methods**: Initial ZSL approaches (Lampert et al. 2013, Jayaraman and Grauman 2014, Changpinyo et al. 2016, Al-Halah et al. 2016) utilized a two-step process involving attributes to classify images from unseen categories. These methods relied on predefined attribute descriptions that could be shared across categories.

2. **Direct Embedding Methods**: More recent research (Frome et al. 2013, Romera-Paredes and Torr 2015, Akata et al. 2015b a, Kodirov et al. 2017) has shifted toward directly establishing connections between image features and semantic spaces, eliminating the need for explicit attribute definitions.

3. **Complex Embedding Approaches**: Many ZSL methods develop sophisticated, multi-faceted embeddings (Socher et al. 2013, Akata et al. 2015a, Xian et al. 2016), primarily aiming to connect image characteristics with semantic space effectively.

4. **Shared Space Alignment**: Another strategy in ZSL involves aligning both image and semantic features into a common intermediary space (Zhang and Saligrama 2015, Fu et al. 2015, Zhang and Saligrama 2016), enabling more direct comparisons between modalities.

In the specific context of SBIR, zero-shot retrieval faces additional challenges due to the cross-modal nature of the task. Zero-Shot SBIR (ZS-SBIR) methods, such as those proposed by Shen et al. (2018), Yelamarthi et al. (2018), and Dey et al. (2019), must bridge not only the category gap but also the modality gap between sketches and images.

2.3.2 Few-Shot Learning for SBIR

Few-shot learning addresses scenarios where only a limited number of labeled examples are available for new categories. This paradigm is particularly relevant for SBIR, as acquiring labeled sketch-photo pairs is resource-intensive.

In practical SBIR settings, datasets typically contain fewer sketches than images for each category, and new datasets often have a limited number of labeled examples. Few-shot learning enables the adaptation of classifiers trained on known datasets to new categories with minimal labeled examples. As an emerging area in transfer learning, few-shot learning has gained significant attention in both machine learning and computer vision (Xian et al. 2019, Schonfeld et al. 2019, Koch et al. 2015, Dong et al. 2021, Wang et al. 2021, Garcia and Bruna 2017).

Few-shot learning approaches in SBIR often build upon zero-shot methods, extending them to scenarios with limited supervision. The key strategies include:

1. **Metric Learning**: Learning a similarity metric that can generalize to new categories with few examples.

2. **Meta-Learning**: Training models to quickly adapt to new tasks using only a few examples.

3. **Data Augmentation**: Generating synthetic examples to expand the limited training data for new categories.

4. **Transfer Learning**: Leveraging knowledge from related categories to improve performance on target categories.

From a practical standpoint, evaluating retrieval performance in few-shot scenarios is essential, as these situations closely approximate real-world applications where complete datasets for all categories are rarely available.

2.4 Domain Adaptation

Domain adaptation addresses the challenge of transferring knowledge from a source domain with abundant labeled data to a target domain with limited or no labeled data. This section explores domain adaptation techniques relevant to SBIR, focusing on methods that can bridge the gap between sketches and images.

2.4.1 Fundamentals of Domain Adaptation

In many application areas, large quantities of unlabeled data are regularly produced, while data labeling remains costly. Domain adaptation (DA) has emerged as a solution to this challenge, enabling the utilization of knowledge from related domains to improve performance on target domains.

The fundamental assumption in traditional machine learning is that training and test data share similar distributions. However, in practice, data distributions often shift across domains or over time (Quiñonero-Candela et al. 2008). These shifts can arise from various factors, such as different lighting conditions, backgrounds, or viewing angles in image-based tasks.

Domain adaptation specifically addresses settings where:

1. Source and target domains have different distributions
2. Source and target domains share the same task (e.g., classifying the same set of categories)
3. The source domain has abundant labeled data, while the target domain has limited or no labeled data

Based on the availability of labeled data in the target domain, domain adaptation approaches can be categorized into:

1. **Unsupervised Domain Adaptation**: No labeled data is available in the target domain (Gong et al. 2012)
2. **Supervised Domain Adaptation**: Labeled data is available in both source and target domains

3. **Semi-supervised Domain Adaptation**: Labeled data from the source domain and a small amount of labeled data from the target domain are available (Mehrkanoon and Suykens 2017)

2.4.2 Shallow Architectures for Domain Adaptation

Domain adaptation problems can be categorized as homogeneous (where source and target domains share the same feature space) or heterogeneous (where the feature spaces differ). For homogeneous domain adaptation, if $\mathcal{X}_s$ represents data from the source domain and $\mathcal{X}_t$ represents data from the target domain, then $\mathcal{X}_s = \mathcal{X}_t$ but $P(\mathcal{X}_s) \neq P(\mathcal{X}_t)$.

Several shallow architecture techniques have been developed for domain adaptation:

Instance Re-weighting Methods: These techniques assign different weights to each training instance to reduce the discrepancy between source and target distributions (Sugiyama et al. 2007). Common approaches include calculating weights based on the density ratio between domains, jointly optimizing weights and classifier parameters (Chu et al. 2013), or using maximum entropy principles to determine optimal re-sampling weights (Shimodaira 2000).

Feature Transformation: A central challenge in domain adaptation is developing feature representations that work effectively for both source and target domains. Transfer Component Analysis (TCA) (Pan et al. 2010) exemplifies this approach by identifying shared latent features between domains while ensuring similar distributions and preserving local geometric structure through a smoothness term.

2.4.3 Deep Architectures for Domain Adaptation

Recent advances in image classification have been driven by deep convolutional architectures trained on large-scale labeled datasets, particularly subsets of ImageNet (Deng et al. 2009). These models not only achieve superior classification accuracy but also produce features that can be repurposed for new tasks (Donahue et al. 2014), even when these tasks differ substantially from the original training objective.

In domain adaptation, initial approaches that simply applied deep features without explicit adaptation already outperformed traditional methods on benchmark datasets like Office (OFF31) (Saenko et al. 2010) and Office+Caltech (OC10) (Gong et al. 2012). Deep Convolutional Activation Features (DeCAF) (Donahue et al. 2014) demonstrated superior performance even without target domain adaptation compared to methods using SURF-BOV features with domain adaptation (Chopra et al. 2013, Donahue et al. 2014, Sun et al. 2016, Csurka et al. 2016).

As shown by Bengio et al. (2013) and Yosinski et al. (2014), deep neural networks learn more abstract and robust representations that encode category-level information and reduce domain-specific biases (Donahue et al. 2014, Sun et al. 2016, Csurka et al. 2016, Saxena and Verbeek 2016).

Deep architectures for domain adaptation can be broadly categorized into three approaches:

1. **Feature Extraction**: Using deep networks as feature extractors and applying traditional domain adaptation techniques to these features. Methods like Feature Augmentation (Daumé III 2009), Max-Margin Domain Transforms (Hoffman et al. 2013), and Geodesic Flow Kernel (Gong et al. 2012) have been applied to DeCAF features (Donahue et al. 2014).

2. **Fine-tuning**: Adapting pre-trained deep networks for new target tasks or domains. This approach typically involves training the network on source domain data and, if available, a small amount of labeled target domain data (Zeiler and Fergus 2014, Oquab et al. 2014, Babenko et al. 2014, Chu et al. 2016).

3. **Domain-Specific Architectures**: Developing deep learning architectures specifically designed for domain adaptation. These include methods for feature adaptation (Glorot et al. 2011), correlation subspace learning (Pan et al. 2010), and cross-domain feature representation (Leggetter and Woodland 1995, Reynolds et al. 2000).

2.4.4 Multi-Source Domain Adaptation

While many domain adaptation methods focus on single-source scenarios, practical applications often involve multiple source domains. Multi-source domain adaptation (MSDA) addresses settings where knowledge from multiple source domains must be transferred to a target domain.

Several approaches have been developed for MSDA:

1. **Feature Augmentation (FA)**: Adding domain-specific feature sets to data representations, with each set originating from a different source domain (Daumé III 2009).

2. **Adaptive Support Vector Machine (A-SVM)**: Utilizing a collection of auxiliary classifiers, each focused on a specific source domain, to fine-tune the parameters of the target classifier (Yang et al. 2007).

3. **Domain Adaptation Machine**: Employing a set of source classifiers along with a domain-related regularizer based on the principle of consistency (Duan et al. 2012).

4. **CP-MSDA**: Assigning weights to each source classifier based on their conditional distributions (Chattopadhyay et al. 2012).

5. **Domain-Specific Class Means (DSCM)**: Relying on domain-specific class means for both metric learning and target class label prediction (Csurka et al. 2015).

6. **Expanded MSDA**: Aggregating domain and classifier regularization terms specific to each source domain (Csurka et al. 2016).

7. **Robust Domain Adaptation via Low-Rank Reconstruction (RDALRR)**: Transforming each source domain into a new representation that allows for linear reconstruction using target domain examples, maintaining connections between reconstructed examples through a low-rank method and identifying outliers through sparsity constraints.

2.5 Semi-Supervised Learning and Domain Adaptation

Semi-supervised learning (SSL) and semi-supervised domain adaptation (SSDA) provide frameworks for leveraging both labeled and unlabeled data, which is particularly relevant for SBIR where labeled data may be limited.

2.5.1 Semi-Supervised Learning

Similar to domain adaptation, semi-supervised learning addresses scenarios with limited labeled data. However, while domain adaptation deals with data from different domains with distributional shifts, SSL focuses on labeled and unlabeled data from the same domain.

With the emergence of deep learning, novel approaches to deep SSL have been developed:

1. **Consistency Regularization**: Ensuring model consistency by comparing predictions on original and augmented data (Laine and Aila 2016, Tarvainen and Valpola 2017).

2. **Virtual Adversarial Training**: Identifying minimal perturbations that destabilize model predictions (Miyato et al. 2018).

3. **Mean Teacher**: Averaging model predictions during training (Laine and Aila 2016) or combining model parameters (Tarvainen and Valpola 2017).

4. **Self-Teaching**: Employing memory modules or monitoring convergence speed for self-supervised learning (Chen et al. 2018, Cicek et al. 2018).

5. **Distribution Alignment**: Directly addressing mismatched data distributions in SSL through augmentation distribution alignment (Wang et al. 2019b).

2.5.2 Semi-Supervised Domain Adaptation

Semi-supervised domain adaptation (SSDA) combines elements of both SSL and domain adaptation, utilizing a structured source distribution along with some labeled samples from the target distribution.

Several approaches have been developed for SSDA:

1. **Enhanced Constraints**: Addressing domain disparities by enhancing constraints on existing labeled data (Donahue et al. 2013).

2. **Subspace Alignment**: Reducing distributional differences by establishing a new subspace (Yao et al. 2015).

3. **Soft Labeling**: Generating soft labels for labeled target samples based on the source model and merging them with hard labels to guide the target model (Ao et al. 2017).

4. **Minimax Entropy Training**: Bridging the gap between unlabeled target samples and class prototypes through entropy-based minimax training (Saito et al. 2019).

2.6 Feature Extraction and Representation Learning

The success of deep learning in computer vision tasks, including SBIR, is largely attributable to its ability to learn complex features directly from raw image pixels. This section examines feature extraction techniques relevant to SBIR.

2.6.1 Deep Convolutional Networks for Feature Extraction

Following the success of AlexNet in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al. 2012), deep convolutional neural networks (CNNs) have become the dominant approach for visual feature extraction. These networks comprise multiple layers, with each layer passing its output to the next in a hierarchical fashion.

Standard CNN architectures include:

1. Linear filters (convolution) followed by non-linear activation functions (e.g., ReLU)

2. Spatial pooling functions (e.g., max/average pooling) for dimensionality reduction

3. Normalization functions (e.g., LRN, batch normalization) for regulating activations

4. Fully-connected layers for final feature representation or classification

5. Loss functions for guiding the learning process through backpropagation

CNN models typically contain millions of parameters, with architectures like AlexNet featuring 60M parameters (Krizhevsky et al. 2012), GoogleNet 4M (Szegedy et al. 2015), and VGG 138M (Simonyan and Zisserman 2014). To prevent overfitting, techniques such as data augmentation, weight regularization, and dropout are employed.

During training, CNN layers learn to respond to specific input patterns, forming a hierarchical representation of features. As demonstrated by Zeiler and Fergus (2014) in their DeconvolutionNet study, lower layers capture basic features like color and edges, while upper layers recognize more complex patterns like eyes, wheels, and faces.

2.6.2 Feature Representation for Image Retrieval

Deep CNN models have been adapted for Content-Based Image Retrieval (CBIR) by utilizing activations from upper layers as image descriptors (Sharif Razavian et al. 2014, Babenko et al. 2014). These features can be further refined through similarity learning or fine-tuning on target datasets, as shown by Wan et al. (2014).

Several techniques have been developed to enhance feature representations for retrieval:

1. **Regional Maximum Activation of Convolutions (R-MAC)**: Combining max-pooled activations from various image regions at different scales across CNN layers (Tolias et al. 2015).

2. **Contrastive and Triplet Losses**: Improving image representations through similarity-based learning objectives (Gordo et al. 2016, Hoffer and Ailon 2015, Radenović et al. 2016, Schroff et al. 2015, Wang et al. 2014, Wang and Gupta 2015).

3. **Multi-scale Orderless Pooling (MOP-CNN)**: Extracting CNN features at various scales and combining them using Vector of Locally Aggregated Descriptors (VLAD) (Gong et al. 2014).

4. **Deep Hashing**: Converting deep features into binary forms for efficient retrieval, either through supervised approaches (Lin et al. 2015, Gao et al. 2015) or unsupervised methods focusing on quantization loss, code balance, and rotation invariance (Lin et al. 2016).

5. **Context-Based Learning**: Using spatial context to train models for predicting the relative positions of image patches (Doersch et al. 2015).

2.7 SBIR Datasets and Evaluation Protocols

The development and evaluation of SBIR methods rely on several benchmark datasets, each with its unique characteristics and challenges.

2.7.1 Benchmark Datasets for SBIR

Sketchy Dataset: Originally introduced by Li et al. (2017), this dataset contains 75,471 hand-drawn sketches of 12,500 objects across 125 categories. It was expanded by Liu et al. (2017) with 60,502 real images from ImageNet, resulting in an average of 484 images per category. The Sketchy dataset is known for its relatively detailed and less abstract sketches, which, while comprehensive, can present challenges in practical applications due to their relative lack of abstraction compared to casual sketches.

TU-Berlin Dataset: Developed by Saavedra (2014), this collection comprises 250 categories of hand-drawn sketches, with approximately 80 sketches per category. It was later augmented by Liu et al. (2017) with 191,067 ImageNet images for SBIR purposes, yielding an average of 764 images per category. The TU-Berlin dataset exhibits greater diversity in sketching styles compared to the Sketchy dataset, with more abstraction and variability.

QuickDraw Extended Dataset: This massive collection contains approximately 50 million drawings across 345 categories, sourced from the Quick, Draw! Game (Wang et al. 2015a). The sketches in this dataset are often more abstract and simplified compared to those in Sketchy and TU-Berlin, reflecting casual drawing styles.

DomainNet Dataset: Encompassing 345 object categories across six domains—clipart, real-world photos, sketches, infographics, paintings, and quickdraw—this dataset (Saavedra et al. 2015) offers a diverse range of visual representations for domain adaptation research.

These datasets often overlap in categories but differ significantly in their characteristics, primarily due to variations in sketching abilities and styles, which introduce different levels of abstraction and uncertainty.

2.7.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) (Hotelling 1992) is a technique for discovering linear projections of two views that maximize correlation while ensuring orthogonality within each view. This method has been extensively used and extended for cross-modal matching tasks, including SBIR.

Key extensions and variants of CCA include:

1. **Regularized CCA**: Incorporating ridge regression to improve stability (Vinod 1976).

2. **Kernel CCA (KCCA)**: Applying non-linear transformations through kernel functions (Akaho 2006, Melzer et al. 2001, Bach and Jordan 2002, Hardoon et al. 2004).

3. **Scalable KCCA**: Employing random Fourier features (FKCCA) or Nyström approximation (NKCCA) for computational efficiency (Lopez-Paz et al. 2014).

4. **Deep CCA (DCCA)**: Modeling transformation functions using deep neural networks (Andrew et al. 2013).

5. **CorrNet**: An encoder-decoder architecture that maximizes correlation between view projections without computing canonical components (Chandar et al. 2016).

6. **Deep Canonically-correlated Autoencoder (DCCAE)**: An encoder-decoder model optimizing both CCA formulation and input reconstruction (Wang et al. 2015b).

7. **Non-parametric CCA (NCCA)**: A non-parametric approach achieving performance comparable to DCCA without neural networks (Michaeli et al. 2016).

8. **Soft-CCA**: Replacing strict decorrelation requirements with more flexible constraints for improved efficiency and scalability (Chang et al. 2018).

9. **Soft-HGR**: A neural framework optimizing a softer formulation of Hirschfeld-Gebelein-Reónyi maximal correlation (Wang et al. 2019a, Hirschfeld 1935, Gebelein 1941, Rényi 1959).

10. **ℓ_0-CCA**: Creating a sparse version of DCCA through stochastic gates performing multiplication on input data (Lindenbaum et al. 2021).

In cross-modality retrieval, CCA-based techniques have been employed to build shared embedding spaces capturing the strongest connections between modalities. Notable approaches include Deep CCA for cross-modal retrieval (Yan and Mikolajczyk 2015) and Ranking-CCA, an end-to-end model minimizing pairwise ranking loss for superior retrieval results (Dorfer et al. 2018).

2.8 Summary

This chapter has provided a comprehensive review of literature relevant to the generalization metrics, sketch-based image retrieval, and domain adaptation techniques that form the foundation of this thesis. The exploration began with generalization in deep learning, which is crucial for selecting optimal foundation models. It then examined the evolution of SBIR methods, from traditional approaches to deep learning-based techniques, highlighting the challenges of zero-shot and few-shot learning scenarios. Finally, it delved into domain adaptation techniques, including single-source and multi-source approaches, which are essential for bridging the gap between sketches and images.

This literature review establishes the theoretical groundwork for the methodologies developed in subsequent chapters. Chapter 3 will build upon the generalization concepts explored here to develop an empirical generalization metric for deep networks. Chapters 4 and 5 will leverage the insights from SBIR and domain adaptation literature to propose novel approaches for single-source and multi-source domain adaptation in SBIR, respectively.

Chapter 3

An Empirical Generalization Framework for Model Selection in SBIR

The selection of appropriate deep learning models for Sketch-Based Image Retrieval (SBIR) requires a robust understanding of their generalization capabilities. While significant theoretical progress has been made in establishing generalization error bounds for deep networks, practical applications—such as SBIR—demand intuitive and empirical metrics that allow developers to compare models effectively. This chapter addresses this need by introducing a novel framework for evaluating generalization in deep networks, with particular emphasis on its application to model selection for SBIR systems.

The concept of generalization—a model's ability to perform effectively on new, unseen data—is fundamental to machine learning theory. A model with strong generalization capacity does not merely memorize its training data but extracts underlying patterns that remain valid when confronted with novel examples. This capacity is particularly crucial for SBIR systems, which must bridge the inherent domain gap between sketches and images while handling diverse sketch styles and previously unseen object categories.

Traditional approaches to measuring generalization have relied on theoretical constructs such as VC dimension and Rademacher complexity. However, as reviewed in Chapter 2, these classical measures often prove inadequate for deep neural networks, producing bounds that are too loose to provide practical guidance. Recent research has attempted to address this limitation, with notable contributions from Neyshabur et al. (2018), who developed tighter generalization bounds for two-layer ReLU networks, and Dziugaite and Roy (2017), who introduced methods for calculating non-vacuous PAC-Bayes generalization bounds. Despite these advances, the application of these theories to multilayer networks frequently results in estimates significantly larger than the actual parameter count, highlighting their limited practical utility.

The Predicting Generalization in Deep Learning (PGDL) competition at NeurIPS 2020 (Jiang et al. 2020), building on work by Jiang et al. (2018), attempted to address these limitations. However, comprehensive studies involving extensive hyperparameter searches (Jiang et al. 2019) have shown that current generalization bounds remain ineffective, with the fundamental mechanisms underlying generalization in deep networks still elusive.

While theoretical efforts continue, there is growing recognition of the need for empirical metrics that can provide practical guidance for model selection and optimization. This need is particularly acute in the context of SBIR, where selecting models with strong generalization capabilities is essential for bridging the domain gap between sketches and images. Despite this requirement, there is a notable scarcity of research on empirical generalization metrics for comparative model assessment.

This chapter addresses this gap by introducing a comprehensive framework for evaluating generalization in deep networks, with specific application to model selection for SBIR. my approach builds upon the insights from deep learning generalization research reviewed in Chapter 2, while addressing the practical requirements of SBIR systems identified in my literature review. The proposed framework provides a quantitative assessment of a model's generalization capacity based on both its classification accuracy and its ability to handle diverse, unseen data—two factors that are crucial for effective SBIR performance.

3.1 Methodology

3.1.1 Rationale for the Proposed Approach

The development of my generalization metric is motivated by several key limitations in existing approaches identified in the literature review. First, as demonstrated by Zhang et al. (2021b) and Neyshabur et al. (2018), conventional generalization bounds often fail to explain the empirical success of overparameterized networks. Second, theoretical measures such as VC dimension and Rademacher complexity, while mathematically elegant, typically yield bounds that are too loose for practical applications (Bartlett and Mendelson 2002, Harvey et al. 2017). Third, recent research by Jiang et al. (2020) and Jiang et al. (2019) has shown that many existing generalization metrics lack predictive power when applied to real-world model selection scenarios.

Drawing on these insights, my approach adopts a fundamentally empirical perspective while incorporating theoretical understanding from both traditional and recent generalization research. Rather than attempting to derive tight theoretical bounds, I propose a metric system that directly measures two key aspects of generalization:

1. **Classification accuracy on unseen data**: This reflects the model's ability to make correct predictions beyond its training distribution—the primary goal of generalization.

2. **Diversity handling capacity**: This captures the model's robustness when confronted with varied, novel inputs—a critical requirement for SBIR systems that must accommodate diverse sketching styles and unseen categories.

This dual focus is particularly relevant for SBIR applications, where models must not only achieve high accuracy but also demonstrate robustness to the inherent variability in sketch inputs. As highlighted by Yu et al. (2016) and Song et al. (2017), the abstraction gap between sketches and images, combined with the diversity of sketching styles, creates a challenging generalization problem that conventional metrics fail to adequately address.

my methodology builds on the linear probe approach introduced by Radford et al. (2021) in the CLIP framework. This choice is motivated by the probe's ability to isolate and evaluate the quality of learned representations without the confounding effects of complex classification models. When a simple linear model achieves high performance using features from a pre-trained network, it indicates that the network has learned meaningful, generalizable representations—a crucial capability for effective SBIR systems.

Furthermore, my framework considers three dimensions that significantly affect generalization performance:

1. **Model size**: Following insights from Neyshabur et al. (2018) and Arora et al. (2018), I examine how parameter count influences generalization capacity.

2. **Randomness**: Building on studies of weight fluctuations and optimization dynamics (Hardt et al. 2016), I incorporate a measure of model robustness to parameter perturbations.

3. **Zero-shot capability**: Inspired by zero-shot learning research (Xian et al. 2018), I assess a model's ability to handle completely novel categories—a critical requirement for practical SBIR applications.

This multidimensional approach provides a more comprehensive assessment of generalization than existing metrics, which typically focus on a single aspect of performance.

3.1.2 Framework Components and Implementation

The proposed metric measures the generalization capacity of a model through accuracy (classification correct or error rates) and the diversity of test data (using the Kappa statistic) across three factors: model size, randomness, and zero-shot capability. my framework consists of two primary components:

1. **Benchmarking Testbed**: Produces raw performance data across different dimensions of generalization. 2. **Empirical Generalization Metric System**: Evaluates and quantifies the model's generalization capacity.

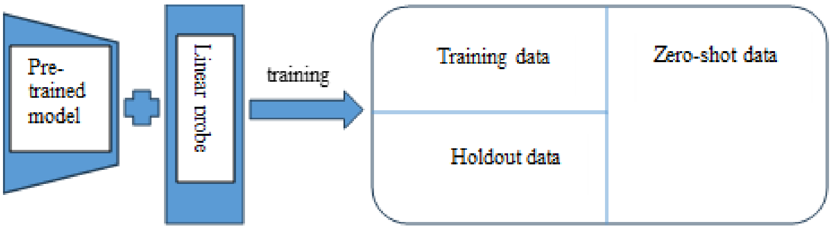


Figure 3.1: (a) Illustration of Benchmark Testbed showing the structure with pre-trained model and linear probe; (b) A 3D array visualization of the generalization metric space, where each cell contains accuracy (g) and diversity (k) measurements. The pink slice represents tests without noise (SSIM=1) and the blue slice represents tests without zero-shot examples.

The Benchmarking Testbed adopts the linear probe structure similar to CLIP (Radford et al. 2021), which provides several advantages for generalization assessment. This approach allows me to evaluate how effectively a deep learning model captures essential features within its hidden layers by training a simple linear model (e.g., logistic regression) using features extracted from a specific layer of the pre-trained network. This methodology is particularly well-suited for generalization assessment because:

1. It isolates representation quality from classifier complexity 2. High performance with a linear probe indicates that the learned features themselves contain the necessary discriminative information 3. It enables fair comparison across different network architectures by standardizing the classification model

As illustrated in Figure 3.1(a), the benchmarking process involves:

1. Pre-training the model on training data 2. Freezing the pre-trained model's weights 3. Training a linear probe on all available data 4. Evaluating the combined model on a mixture of holdout data and zero-shot data

This evaluation process provides insights into both the feature extraction capabilities of the model and its ability to generalize to new or unseen data—two critical factors for SBIR performance.

Following Jiang et al. (2020), I define generalization error as:

$$g\left(f_w; D\right)=\frac{1}{\left|D_{\text{test}}\right|} \sum_{(x, y) \in D_{\text{test}}} \mathbb{1}\left(f_w(x) \neq y\right)-\frac{1}{\left|D_{\text{train}}\right|} \sum_{(x, y) \in D_{\text{train}}} \mathbb{1}\left(f_w(x) \neq y\right) \tag{3.1}$$

where w denotes the model's weight set and $\mathbb{1}$ is the indicator function.

A key innovation in my approach is the systematic investigation of how various hyperparameters affect generalization performance. Different hyperparameter configurations result in diverse weight values, producing many variants of a given model architecture. To capture this variability and its impact on generalization, I draw inspiration from Jiang et al. (2020) and sample weight values across a spectrum of hyperparameter types.

my experimental design includes repeated model training on identical data while monitoring weight fluctuations. As shown in Figure 3.2, these fluctuations typically follow a normal distribution pattern. To account for this inherent randomness, I establish a framework where each weight in a pretrained model defines the center of a sampling window. By taking multiple random samples within these windows (over 50 in my implementation), I generate numerous model variations. The average performance across these variations serves as a robust benchmark for a given window size, with the window size itself indicating the level of randomness being tested.

This approach allows me to systematically explore the relationship between model robustness to parameter perturbations and generalization performance—a relationship that is particularly relevant for SBIR systems, which must be robust to the inherent variability in sketch inputs.

Empirical Generalization metric system is to seek for a trade-off point to illustrate the generalization of test models, as shown in Algorithm 1.

Step 1. Experimentally, I calculate the accuracy of individual classes on test data using Eq.3.1 and assess the diversity of the test data, encompassing both holdout data and zero-shot data. The former computation allows for deriving a distribution of error rates across all classes, whereas generalization error typically pertains to the overall error rate. On the other hand, the

Algorithm 1 Empirical Generalization Metric System

- **Preparation:** training a given model and its invariants, including training linear prob layer.
- Computing accuracy and Kappa by Eq.3.1&Eq.3.6;
- Computing similarity of each cell and the origin cell by KL-D on the accuracy and Kappa respectively, and updating the 3D array with these two kinds of KL-D;
- Searching a trade-off point on the updated 3D array by Eq.3.3&Eq.3.4;

measurement of diversity can be achieved using the Kappa statistic [Cohen, 1960], Given a dataset with n classes, I may divide all the classes into two parts according to the current class i, that is, the i-th class and non i-th class. The classification results can be described as, classifying a sample to the i-th class is denoted as $h_i = 1$ and not to the i-th class as $h_i = -1$; classifying a sample to the group $\{h_j : j \neq i, j = 1..n\}$ is denoted as $h_{j \neq i} = 1$ and not to the group $\{h_j : j \neq i, j = 1 \ldots n\}$ is denoted as $(h_{j \neq i}) = -1$. The confusion matrix is defined as below:

	$h_i = 1$	$h_i = -1$
$\{h_{j \neq i}\} = 1$	#a	#c
$\{h_{j \neq i}\} = -1$	#b	#d

where #a is the number of samples predicted as positive in line with the events h_i and $\{h_{j \neq i}\}$, and similarly for #b, #c, and #d. It is conceivable that certain samples may be unrecognized by the benchmarking model due to excessively high loss or low probability in the model outputs. Thus, I set a threshold to identify such failed samples and count them in ”#d”. Moreover, it is possible that there are samples not to be recognised by the model since the model outputs very close probabilities for multiple candidate classes, including the i-th class. I simply set a threshold for probability difference to identify the conflict cases, i.e. the events of h_i and $\{h_{j \neq i}\}$ agree with each other in the confusion matrix, and count them to "#a". The others count to "#b or #c". The Kappa is defined as,

$$\begin{aligned} k_i &= \frac{p_1 - p_2}{1 - p_2} \\ p_1 &= \frac{a + d}{N} \\ p_2 &= \frac{(a+b)(a+c) + (c+d)(b+d)}{N^2} \end{aligned} \tag{3.2}$$

where N denotes the number of total class samples. A model with strong generalisation capacity should be adaptable to highly diverse data. When Kappa is high, it means that the model classifies different classes and results in the conflict cases too much. The model has a low diversity, i.e. low generalisation capacity. Otherwise, it has a high diversity, i.e. high generalisation capacity.

Step 2. I gather measured data by Eq.3.1 and Eq.3.6 across three distinct dimensions: model size (representing the number of weights), randomness (indicated by the window size), and the proportion of zero-shot data (comprising one hundred percent of the test data). This compilation results in a three-dimensional array as shown in Figure 3.3. Each cell within this array records both the accuracy "g" and Kappa metrics "k". I focus on test data diversity. Thus, I collect the correct rates or error rates of each class as the accuracy data "g" and the Kappa data of each class as the diversity data "k". The resulting "g and k" adhere to distributions across the classes of test data and are stored in one cell. Different cells within the 3D array correspond to their individual settings of three dimensions. Employing a Histogram on a singular cell (denoted as a pair of "g and k") makes its distribution visualized in Figure 3.4.

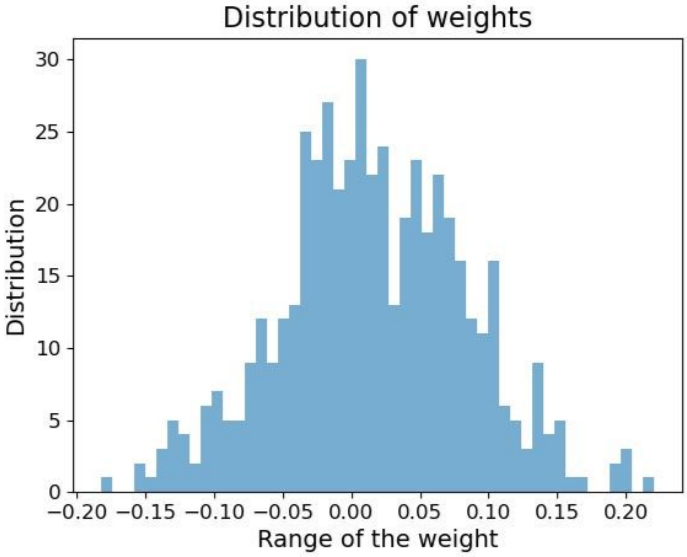


Figure 3.2: fluctuation of a weight with 500 trainings

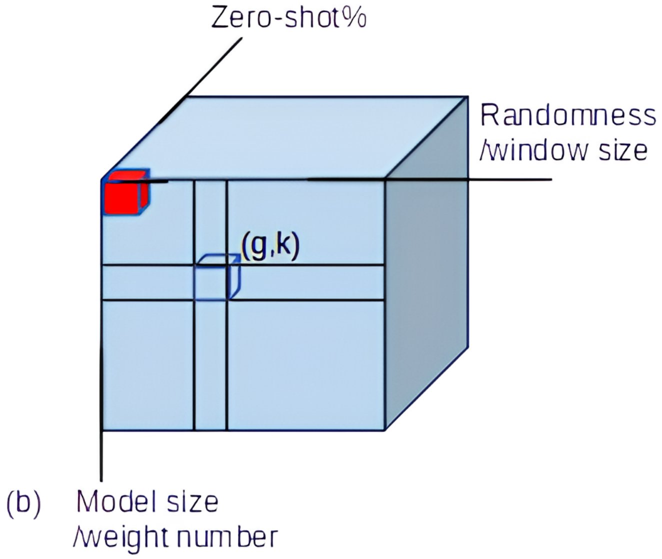


Figure 3.3: 3D array with three dimensions of zero-shot percent, randomness and model size

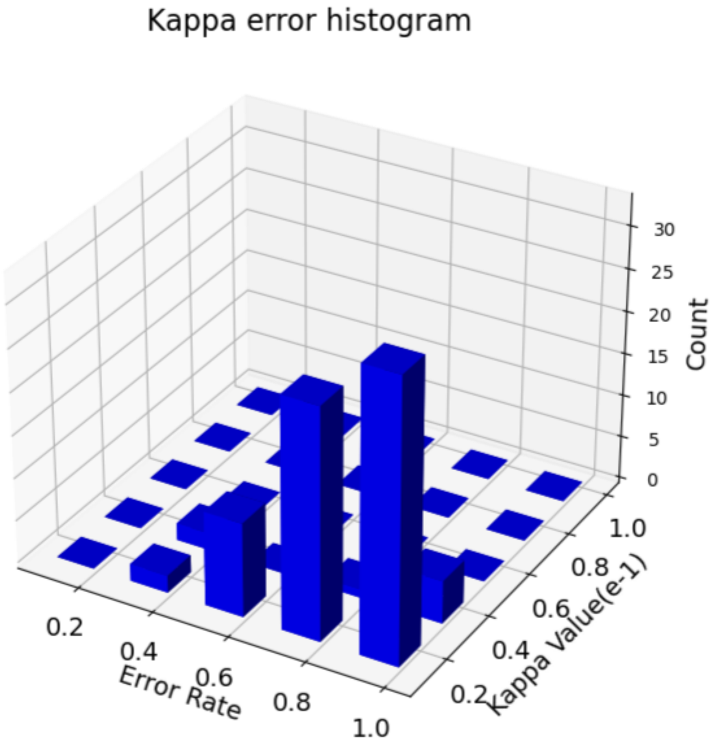


Figure 3.4: Histogram of a singular cell in the figure above

Step 3.I compute the similarity of each cell and the origin cell of the 3D array by KL-Divergence on the accuracy and Kappa respectively, and denote the KL-Divergence of the accuracy as KL_g (ZeroShot, Rand, WeightNum) and the KLDivergence of the Kappa as KL_k (ZeroShot, Rand,WeightNum) here. The origin cell refers to the setting of non-zero-shot data (i.e. holdout data), non-randomness on model, and the model maximum size (see red block in Figure 3.3). For a test model, the origin cell should be of its optimal performance. Ideally, both kinds of KL-Divergences tend to Zero, that is, the model with different settings can still approach the optimal performance. In practice, I used to employ Jensen-Shannon Divergence instead for data visualization purpose.

Step 4. I estimate the trade-off point based on the 3D array updated with the two kinds of KL-Divergences, i.e., KL_g (ZeroShot,Rand,WeightNum) and KL_k (ZeroShot,Rand,WeightNum). Searching the tradeoff point on the 3D array is expressed as,

$$\text{TradeOff } = \arg \min_{(x,y,z)\in 3DA} \| KL_g(x,y,z) - KL_k(x,y,z) \|^2 \tag{3.3}$$

For such multivariate optimization problem, I apply the marginalization approach to the KL_g and KL_k as below,

$$\begin{cases} KL_g(x \sim 3DA(\text{ ZeroShot })) = \sum_{(y,z)\sim 3DA(\text{ Rand,WeightNum })} KL_g(x,y,z) \\ KL_k(x \sim 3DA(\text{ ZeroShot })) = \sum_{(y,z)\sim 3DA(\text{ Rand,WeightNum })} KL_k(x,y,z) \end{cases} \tag{3.4}$$

There are three pairs of marginal distributions in total. Each pair results in a cross point. These 3 cross points indicate the values of three dimensions separately, i.e. model size, randomness, and zero-shot percentage, which is called as the trade-off point.

3.2 Experiment

This section I follow the step in the methodology and record the result under different settings.

3.2.1 Implementation details

I use the CIFAR-100 dataset (Krizhevsky et al. 2009) for training and test, which consists of 60,000 color images associated with 100 classes. Each class contains 600 images. In my experiments, I pick up 50 classes for training and the rest 50 classes for the zero-shot scenario tests. I use the Two-Layer-ReLU network presented in (Neyshabur et al. 2018) and a simple CNN network as shown in Figure 3.5. The model size is indicated by the parameter number of the model in my tests. In the step 0 of Algorithm 1, the models and their invariants (e.g. the same original model with different model sizes) are pre-trained based on the training data. All the data, models, and benchmarking results are available on

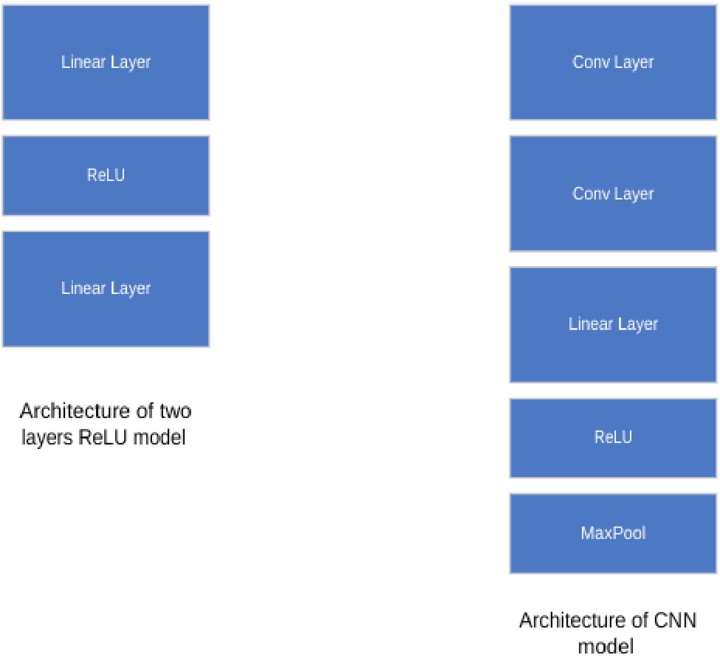


Figure 3.5: two test model architecture

3.2.2 Tests

I organise my experiments in terms of the Algorithm 1 to illustrate how to use the proposed empirical generalisation metric. As there is a lack of similar work available for comparison, I only benchmark two models in Figure 3.5. Alongside the illustration of each step in Algorithm 1, the intuitive observations are given as well. These intuitive and qualitative analysis can be quantified by the trade-off point. Finally, I conclude the benchmarking results through the trade-off points.

Step 1. Collect accuracy and Kappa data of a model Accuracy data.I test the model of Two-Layer-ReLU and store the correct rates of each class in each cell of a 3D array. The average of all the correct rates in one cell represents the accuracy of the model with a specific setting of three dimensions. Figure 3.6 shows the changes of accuracy along with different dimensions. For the dimension of randomness (i.e. window size), I set 5 window sizes for each parameter of the model, i.e. 0,0.2,0.4,0.6,0.8, and sample 50 random values for each parameter at every window size level. Although each data point in Figure 3.6 is the average of 50 trials, randomness still appears by big ups and downs on curves. It can be noted that randomness results in the accuracy decreasing quickly. Moreover, Figure 3.6a shows the results without randomness (i.e. window size is Zero). It can be noted that small models have low accuracy and the accuracy can be increased when model size (i.e. parameter number) increasing. Obviously, over-parameterization stops this trend. It is because when model size exceeds the size of training data, increasing model size continuously will not bring about the accuracy improvement. I further illustrate the performance of accuracy and window-size in Figure 3.7. I select 4 model sizes here, i.e. parameter numbers of 1M, 16M, 32M, 64M. It is clear that randomness results in the accuracy decreasing quickly.

Moreover, Figure 3.6a shows the results without randomness (i.e. window size is Zero). It can be noted that small models have low accuracy and the accuracy can be increased when model size (i.e. parameter number) increasing. Obviously, over-parameterization stops this trend. It is because when model size exceeds the size of training data, increasing model size continuously will not bring about the accuracy improvement. I further illustrate the performance of accuracy and window-size in Figure 3.7. I select 4 model sizes here, i.e. parameter numbers of 1M, 16M, 32M, 64M. It is clear that randomness results in the accuracy decreasing quickly.

Statistic Kappa. I set 5 window sizes for each parameter of the model, i.e. 0,0.2,0.4,0.6,0.8, to obtain 50 random values for each parameter at each window size level. I test the model of Two-Layer-ReLU and store the Kappas of each class in each cell of the 3D array. The average of all the classes' Kappas in one cell represents the diversity of the model with a specific setting of three dimensions. Figure 3.8 shows the change of diversity along with different dimensions. It can be noted that the Kappa is increasing when adding zero-shot data.

Step 2.Histogram. The measure data of accuracy and Kappa, including the correct rates and Kappa values of each class, is cell-wise stored in the 3D array according to three dimensions. Figure 3.6,Figure 3.7 and Figure 3.8 may provide insights into potential performance trends rather than facilitating a

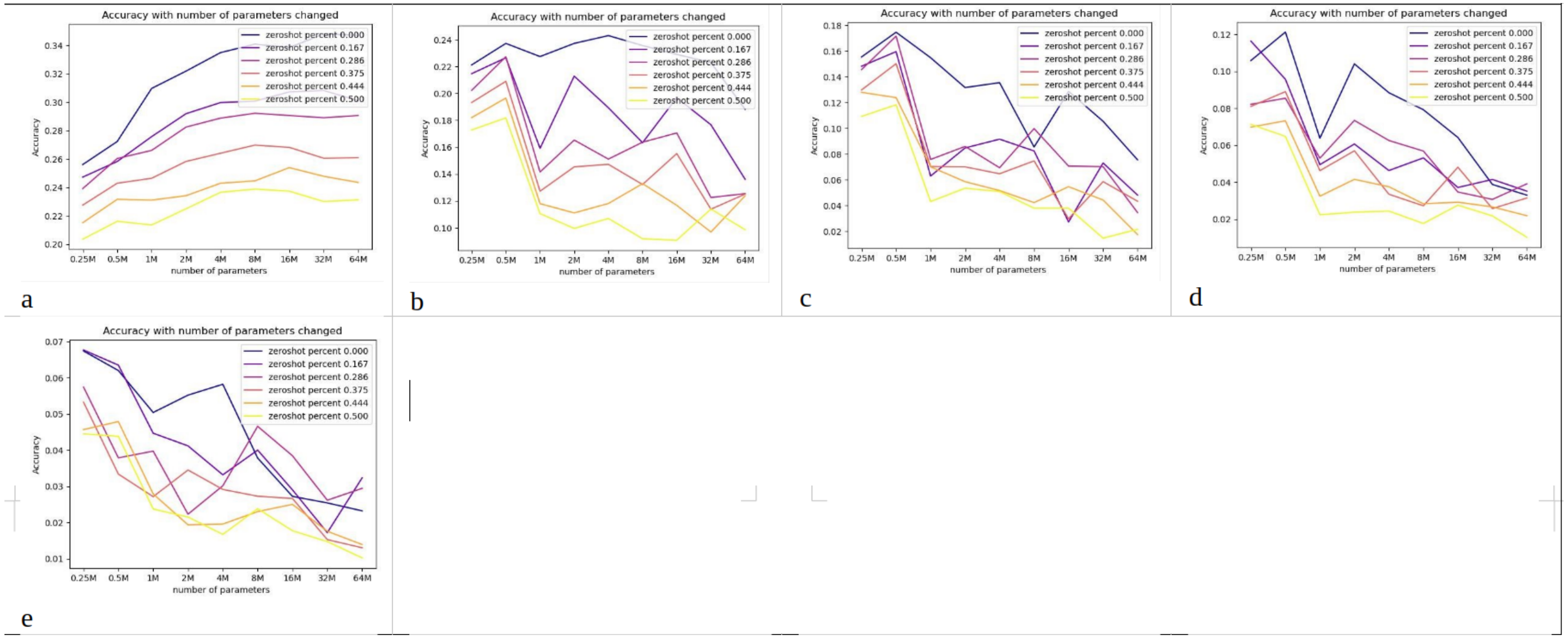


Figure 3.6: Performance of accuracy along with 3 dimensions. (a)-(e) correspond to 5 window sizes separately.

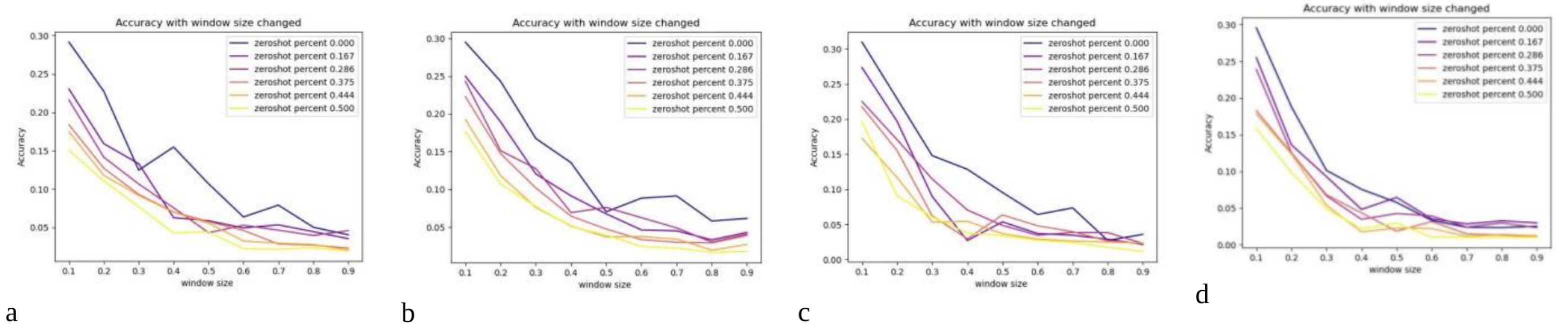


Figure 3.7: Performance of accuracy with randomness dimension (i.e. window size). (a)-(d) correspond to 4 model sizes respectively.

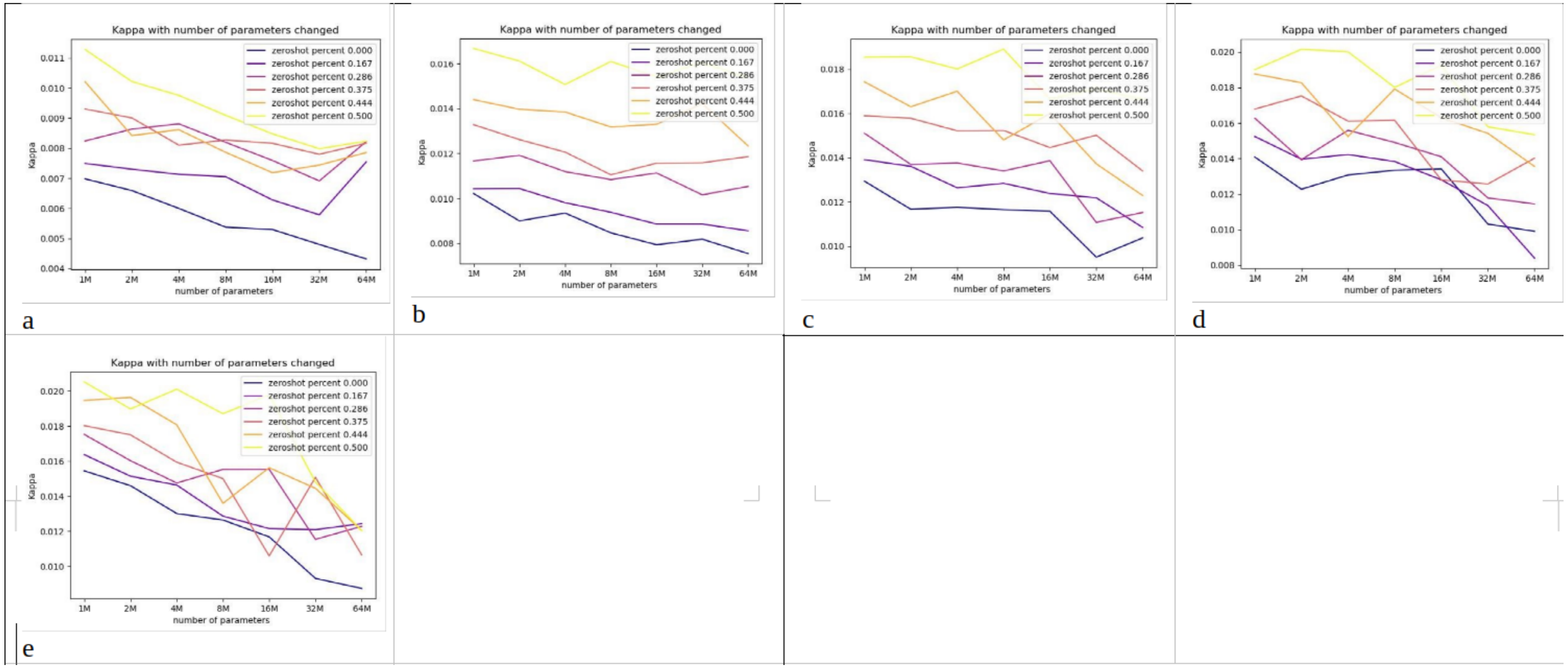


Figure 3.8: Performance of diversity (Kappa) along with 3 dimensions. (a)-(e) correspond to 5 window sizes separately..

quantitative analysis. In fact, each cell contains two distributions—namely, accuracy and diversity distributions. I apply Histogram to this 3D array to illustrate these two distributions in cells. Figure 3.9 shows the Histograms of the original cell (best performance) and of the cell (worst performance) with maximum zero-shot data, maximum randomness and smallest model size in the 3D array.

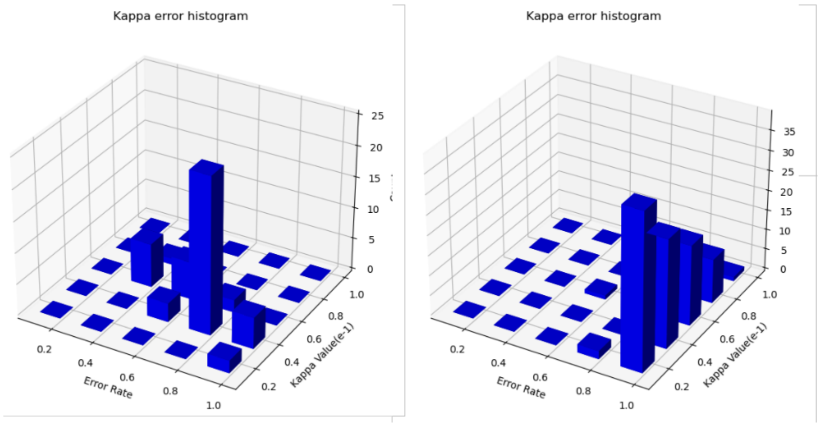


Figure 3.9: Histograms of the best (left) and worst (right) cells.

Step 3.JS-Divergences I apply Jensen-Shannon divergence to the 3D array and update it with the resulting JS divergences. I set 5 window sizes for each parameter of the model here, i.e. $\{0, 0.2, 0.4, 0.6, 0.8\}$. The JS divergence of accuracy is shown in Figure 3.10. It can be noted that the JS-Divergence is increasing along with the window size increasing. This implies that randomness brings about big JS-divergences of accuracy. Moreover, I show the performance of accuracy JS-divergences and randomness in Figure 3.11. I set 6 model sizes here, i.e., parameter numbers of $\{1\text{M}, 2\text{M}, 4\text{M}, 8\text{M}, 16\text{M}, 32\text{M}\ \}$. Figure 3.11 enhances my observation in Figure 3.10, i.e. randomness brings about big divergences. The JS divergence of Kappa is shown in Figure 3.12. The key observation is that when the model size exceeds the training data size, JS-Divergences rapidly decrease and subsequently tend to stabilize or smooth out. This is due to the over-parameterization. Moreover, along with the randomness increasing (i.e. window size) and zero-shot data involving, JS-Divergences are increasing. However, it can be noted that for the non-zero-shot-data scenarios (i.e. zero-shot=0), JS-Divergences have small changes. This implies that this model is not adaptable with zero-shot scenarios. The model's diversity is very limited.

Furthermore, I show the performance of Kappa JS divergence and randomness (i.e. window size) in Figure 3.13 . I select 6 model sizes here, i.e. parameter numbers of 0.25M, 0.5M, 1M, 16M, 32M, 64M. It is obvious that when increasing randomness, all the JS divergences are converging. This implies that randomness results in all the performances converging and decreasing. The model has the low randomness

Step 4.Trade-off point.I compute the trade-off point based on 3 pairs of marginal distributions and show them in Figure **??**a gives the cross point at (0.30, 0.26), Figure **??**b gives the cross point at (0, 0.24), and Figure **??**c gives the cross point at (809832, 0.28). The trade-off point for this model is of (Rand=0.30, ZeroShot=0, WeightNum=809882). The average of marginals is around 0.26, which reflects the overall performance. Moreover, I also test the CNN model and show its trade-off point as well as the trade-off point of Two-Layer-ReLU model in Figure **??**. It can be noted that**??**f has multiple cross points. Usually, I prefer to small size of the model. Thus, the cross point in **??**f is selected at (3634764, 0.33). The trade-off point is of (Rand=0.58, ZeroShot=0.01672616, WeightNum=3634764). The average of marginals is around 0.34. Two trade-off points are shown as below for comparison,

Model	Marginal	Rand	Zero-Shot(%)	WeightXum
TwoLaxer	0.26	0.30	0	809882
CNN	0.34	0.58	0.01673	3634764

It can be noted that the Two-Layer-ReLU model has no transfer learning ability since zero-shot percentage is zero, while the CNN model has certain transfer learning ability against the Two-Layer-ReLU model. However, the model size of CNN model is increasing drastically when generalization capacity increasing. Obviously, a big size of the model is not widely acceptable.

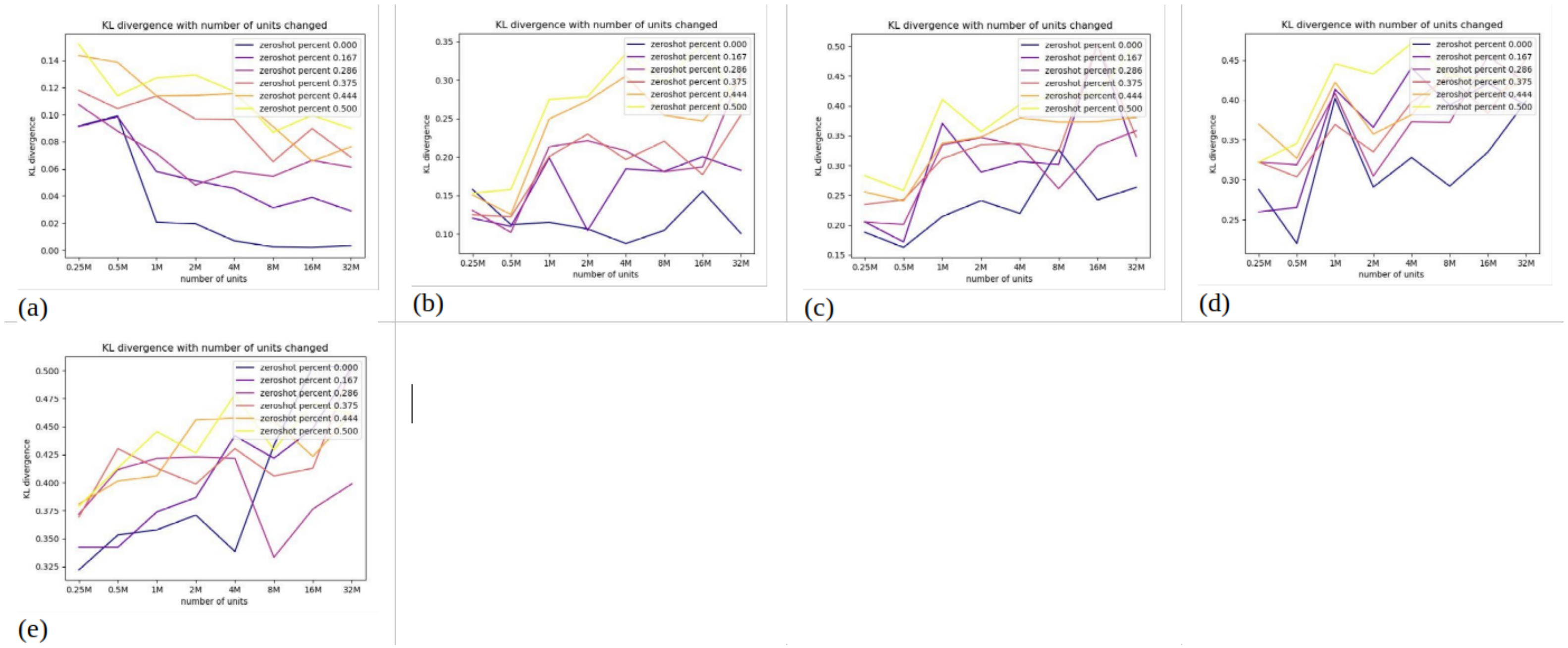
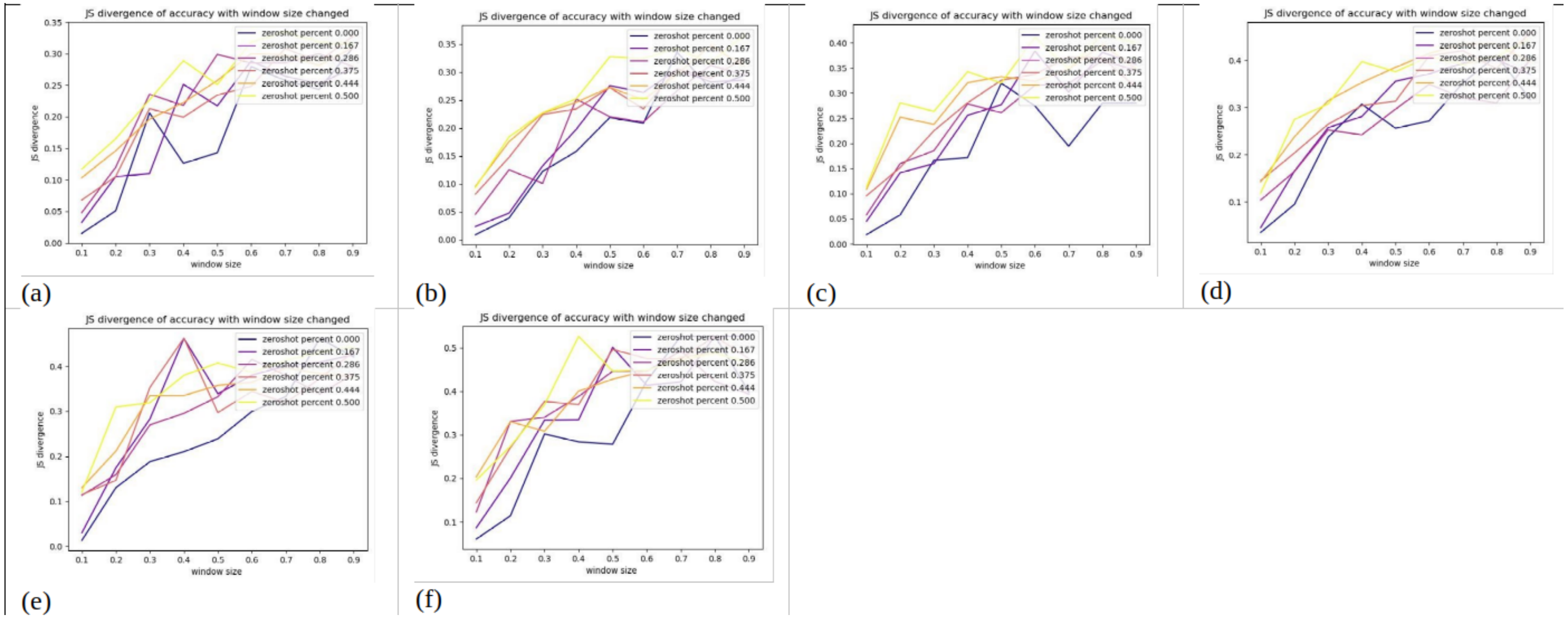


Figure 3.10: Performance of accuracy JS-D along with 3 dimensions. (a)-(e) correspond to 5 window sizes separately.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.11: Performance of accuracy JS-D with randomness dimension. (a)-(f) correspond to 6 model sizes separately.

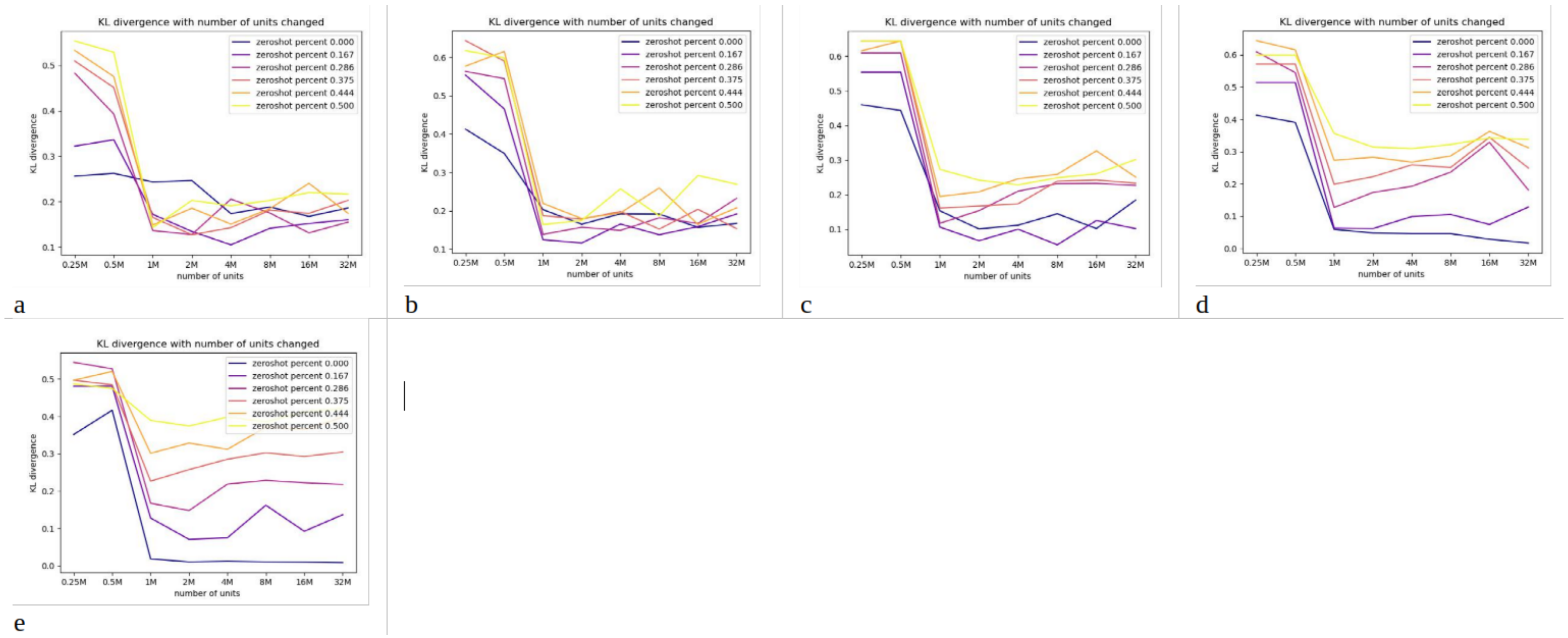


Figure 3.12: Performance of Kappa JS-D along with 3 dimensions. (a)-(e) correspond to 5 window sizes.

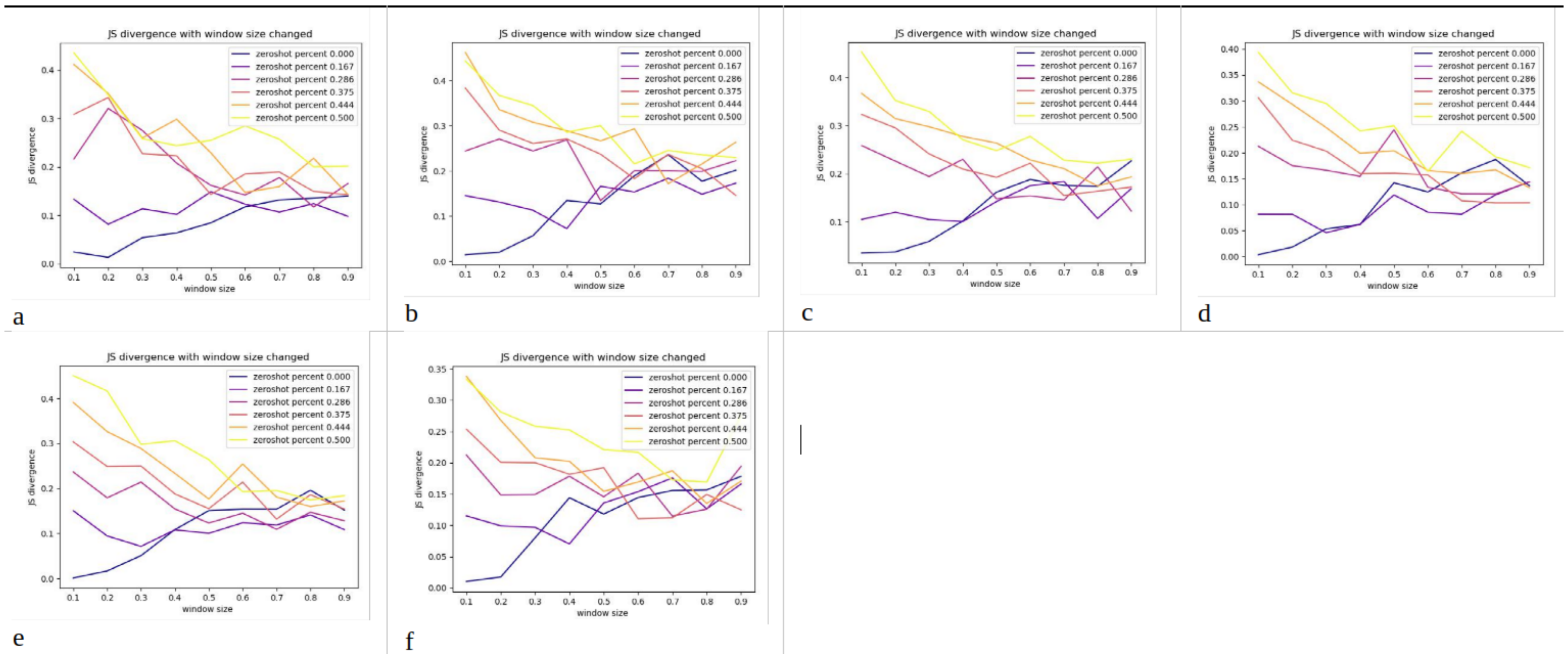


Figure 3.13: Performance of Kappa JS-D with randomness dimension. (a)-(f) correspond to 6 model sizes.

3.3 Experiment on deeper neural network

In this section I updated my benchmark and test it on some deeper neural networks which is widely used by researchers right now.

3.3.1 The updated benchmark design for SOTA deep network

The updated benchmark is still to seek for a trade-off point to illustrate the generalization of the test models, but with a little modification as shown in the following steps:

$$g\left(f_w; D\right)=\frac{1}{\left|D_{\text {test }}\right|} \sum_{(x, y) \in D_{\text {test }}} \mathbb{1}\left(f_w(x) \neq y\right)-\frac{1}{\left|D_{\text {train }}\right|} \sum_{(x, y) \in D_{\text {train }}} \mathbb{1}\left(f_w(x) \neq y\right) \tag{3.5}$$

Step 1.I compute the ErrorRate of individual classes on the test data using Eq.1. It enables the derivation of a distribution of error rates across all classes, while the generalization error typically refers to the overall error rate. I then evaluate the diversity of the test data using the Kappa statistic (Cohen (1960)). In the context of multi-class classification problem, I are dealing with agreement and disagreement among classifier outputs. The Kappa is indeed more robust than simple percentage agreement because it adjusts for the possibility of agreement occurring by chance. This is particularly useful when there is a class imbalance, as chance agreement would be higher for the more frequent classes. Similarly, it also results in a distribution of Kappa across all classes.

Given a dataset with multiple classes, I may divide all the classes into two parts according to the current class i, that is, the i-th class and non i-th classes. The classification event is denoted as $h_i(x) = 1$ for classifying x into the i-th class or $h_i(x) = -1$ for classifying x not into the i-th class. Similarly, $h_{\bar{i}}(x) = 1$ for classifying x into the non i-th classes or $h_{\bar{i}}(x) = -1$ for classifying x not into the non i-th classes. The classification results can be described as $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $y_i \in \{-1, 1\}$ are the class labels of binary classification. The confusion matrix of the $\{h_i\}$ and $\{h_{\bar{i}}\}$ for binary classification is

	$h_i = 1$	$h_i = -1$
$h_{\bar{i}} = 1$	$\#a$	$\#c$
$h_{\bar{i}} = -1$	$\#b$	$\#d$

where $\#a$ represents the number of samples predicted as positive in line with the events h_i and $h_{\bar{i}}$, and similarly for $\#b, \#c, \#d$. For example, when the fine-tuned model outputs high but very close probabilities for multiple candidate classes, including the i-th class, this results in conflict. The samples can not be recognized by the model. I thus count them in "$\#a$". When the fine-tuned model outputs low but very close probabilities for multiple candidate classes, including the i-th class, this results in conflict as well. The samples cannot also be recognized by the model. I thus count them in "$\#d$". It can be noted that $\#a$ and $\#d$ refer to conflict case numbers while $\#b, \#c$ refer to conflict-free case numbers. It is conceivable that certain samples may go unnoticed by the fine-tuned model due to excessively high loss or low probability in the model outputs. Therefore, I set a threshold to identify such failed samples and count them in "$\#d$". The Kappa about the i-th class is defined as,

$$\begin{cases} k_i = \dfrac{p_1 - p_2}{1 - p_2} \\ p_1 = \dfrac{a+d}{N}, p_2 = \dfrac{(a+b)(a+c) + (c+d)(b+d)}{N^2} \end{cases} \tag{3.6}$$

where N denotes the number of total class samples. The average of the Kappas for all the classes may be regarded as the generalization Kappa. A model with strong generalization capacity should be adaptable to highly diverse data. When the Kappa statistic is high, it indicates that the model is struggling to properly classify samples into different classes, leading to an excessive number of conflict cases. This suggests that the model has low diversity, and consequently, a low generalization capacity. Conversely, if the Kappa statistic is low, it implies that the model exhibits high diversity, and therefore has a high generalization capacity.

Different from the previous method, I take the robustness into consideration and remove the randomness dimension from my test bed. Regarding the robustness dimension, in deep learning, robustness measures how well a network performs under controlled variations such as noise or distortions, providing insights into the network's ability to generalize effectively (Natekar and Sharma (2020)). This concept is extended to adversarial robust learning settings under the umbrella of adversarial robustness. Recent works focus on the generalization gap in robust learning contexts (Zhang et al. (2021a), Yang et al. (2020)). Further exploration of robust generalization challenges in adversarial learning models can be found in (Li et al. (2022) and Kim et al. (2023)). Moreover, (Bubeck and Sellke (2023)) highlights that "over-parameterization" is also necessary for robust learning. Consequently, robustness is incorporated into my testbed by introducing adversarial samples into the test data.

Within the three dimensions (zero-shot%, weight number, robustness) of the 3D array, I can calculate two distributions on a cell-wise basis: one related to ErrorRate and the other to Kappa. These calculations are carried out by Eq.3.1 for ErrorRate and Eq.3.6 for Kappa, and are stored within the 3D array (denoted as a pair of "g and k" for each cell, see Figure3.1b).

I depict these two distributions of each cell by three kinds of statistics, i.e., means (denoted as M), standard deviations (denoted as SD), and $10th$ percentiles (denoted as ${}^{10}P$). The $10th$ percentile score indicates that 10% of the trials scored below it. Since smaller means are better in this context, the $10th$ percentiles represent the best performing 10% of classification outcomes.

I update each cell in the 3D array by these three kinds of statistics with respect to two distributions (i.e., ErrorRate and Kappa) within three dimensions, that is, $M_g(ZeroShot, Robust, WeightNum)$, $SD_g(ZeroShot, Robust, WeightNum)$, ${}^{10}P_g(ZeroShot, Robust, WeightNum)$ on ErrorRate and $M_k(ZeroShot, Robust, WeightNum)$, $SD_k(ZeroShot, Robust, WeightNum)$, ${}^{10}P_k(ZeroShot, Robust, WeightNum)$ on Kappa.

Step 3. I estimate the trade-off point based on the three kinds of statistics within three dimensions in the 3D array. The desired generalization capacity should be achieving high performance of accuracy and diversity by

maximizing two dimensions of zero-shot capabilities and robustness, while minimizing the dimension of model size as much as possible.
Searching the trade-off point over the 3D array ($3DA$) is described as,

$$\begin{aligned} \min_{(x,y,z)\in 3DA} & \big(M_g(x,y,z)+SD_g(x,y,z)+{}^{10}P_g(x,y,z) \\ & +M_k(x,y,z)+SD_k(x,y,z)+{}^{10}P_k(x,y,z)\big) \\ \text{subject to} & \begin{cases} c_1 : x \geqslant ZeroShot_{\min} \\ c_2 : y \geqslant Robust_{\min} \\ c_3 : z \leqslant WeightNum_{\max} \end{cases} \end{aligned} \tag{3.7}$$

where $(ZeroShot_{\min}, Robust_{\min}, WeightNum_{\max})$ are the given maximum(/minimum) bounds of three dimensions. Particularly, I prefer to maximize (or minimize) these bounds for generalization purpose here. Equation3.7 may be converted to a minmax optimization problem as follows,

$$\begin{aligned} \min_{(c_1,c_2,c_3)} \|C\|^2 \\ \text{subject to:} \end{aligned} \begin{cases} \min\limits_{(x,y,z)\in 3DA}\Big(M_g(x,y,z)+SD_g(x,y,z)+{}^{10}P_g(x,y,z)+ \\ \quad M_k(x,y,z)+SD_k(x,y,z)+{}^{10}P_k(x,y,z)\Big) \\ c_1 \geqslant 1-x \\ c_2 \geqslant y \\ c_3 \geqslant z \end{cases} \tag{3.8}$$

where $C = (c_1, c_2, c_3)$ denotes the upper bounds. I apply GEKKO? to minimize the upper bounds of three dimensions (i.e., ZeroShot, Robust, WeightNum) to approach the trade-off point. Ideally, the resulting (x, y, z) would be equal to the resulting (c_1, c_2, c_3). I always select the resulting (x, y, z) as the trade-off point in practice.

To visualize it, I compute the marginal distributions with respect to three dimensions separately. The marginal distributions with respect to the dimension of $ZeroShot$ is computed as,

$$\begin{cases} M_g(x \sim 3DA(ZeroShot)) = \\ \sum_{(y,z)\sim 3DA(Robust,WeightNum)} M_g(x,y,z) \\ SD_g(x \sim 3DA(ZeroShot)) = \\ \sum_{(y,z)\sim 3DA(Robust,WeightNum)} SD_g(x,y,z) \\ {}^{10}P_g(x \sim 3DA(ZeroShot)) = \\ \sum_{(y,z)\sim 3DA(Robust,WeightNum)} {}^{10}P_g(x,y,z) \\ M_k(x \sim 3DA(ZeroShot)) = \\ \sum_{(y,z)\sim 3DA(Robust,WeightNum)} M_k(x,y,z) \\ SD_k(x \sim 3DA(ZeroShot)) = \\ \sum_{(y,z)\sim 3DA(Robust,WeightNum)} SD_k(x,y,z) \\ {}^{10}P_k(x \sim 3DA(ZeroShot)) = \\ \sum_{(y,z)\sim 3DA(Robust,WeightNum)} {}^{10}P_k(x,y,z) \end{cases} \tag{3.9}$$

There are a total of three sets of marginal distributions separately for three dimensions. Each set illustrates the generalization bounds (referred to as $M_g, SD_g, {}^{10}P_g$) and diversity (referred to as $M_k, SD_k, {}^{10}P_k$) concerning the scale at each dimension specified by the trade-off point, one after another. Theoretical equivalence is expected among these three sets of marginal probabilities at the trade-off point.

In fact, the trade-off point indicates the model's tolerance on three dimensions at an expected marginal probability level. The area delimited by the trade-off point intuitively and quantitatively illustrates the generalization capacity of the test model.

3.3.2 Data and Models

I use CIFAR-100 (Krizhevsky et al. (2009)) and ImageNet datasets (Russakovsky et al. 2015) for fine-tuning and tests. In my experiments, I pick up 50 classes for training and the rest 50 classes for the zero-shot scenario tests from CIFAR-100. I randomly select 100 object classes from ImageNet. Similarly, I divide it into two parts, i.e., 50 classes for training and the other 50 classes for tests. These two datasets are widely used in deep learning applications. The primary difference is the image size; ImageNet images are larger than those in CIFAR-100. Larger images in ImageNet provide more data, which generally leads to better learning outcomes. In contrast, the smaller images in CIFAR-100 often result in ambiguity, where additional context is necessary to accurately interpret the images. In addition, I apply augmentation approaches to these datasets to generate unseen data or classes in case that the pretrained models have seen data in their previous training.

I select the CLIP and EfficientNet models for benchmarking tests since they both share similar architecture. They have some connections as well as differences. I use 5 pre-trained CLIP models from Radford et al. (2021) and 8 EfficientNet models from Tan and Le (2019). Table 3.1 shows the pre-trained model sizes of CLIP and EfficientNet respectively. Although these pre-trained models have been optimised, they still need to be fine-tuned with the linear probe on the training data in advance. I only use the weight number of each model as the dimension of model size in the experiments, neglecting the other issues such as layers, depth, the change of structure, so that the pre-trained models line up in an ”over-parameterization” way. I hope to have an insight to the generalisation capacity of these two kinds of pre-trained models, i.e. CLIP group and EfficientNet group. Moreover, the test data is added noises for robustness tests. To quantify noise levels, I employ the Autoencoder to the test data to generate noisy data and use the Structural SIMilarity (SSIM) Index metric to control noise levels. When SSIM is decreasing towards zero, the noise level is increasing. All the experiments work on a Workstation with Nvidia 12G RTX2080. All the data, models, and benchmarking results are available on GitHub.

3.3.3 Trade-Off points of CLIP and EfficientNet

The pre-trained CLIP models (i.e. RNxxx) and EfficientNet models are CNN-based (see Table3.1). For comparison, the CLIP ViT-xxx models are not taken into account here.

EfficientNet	# Params	CLIP	# Params
efficientnet-b0	5.3M	RN50	38M
efficientnet-b1	7.8M	RN101	56M
efficientnet-b2	9.2M	RN50x4	87M
efficientnet-b3	12M	RN50x16	167M
efficientnet-b4	19M	RN50x64	420M
efficientnet-b5	30M	ViT-B/32	87M
efficientnet-b6	43M	ViT-B/16	86M
efficientnet-b7	66M	ViT-L/14	304M

Table 3.1: Pretrained Models' Parameters

Step 1. Collect ErrorRate and Kappa data of both kinds of test models
I test the pretrained models of CLIP and EfficientNet on test data across three dimensions (i.e., zero-shot%, weight number, SSIM) and store the error rates and Kappas for each class in each cell of a 3D array.

Step 2. Update 3D Array
I compute three kinds of statistics related to the distributions of ErrorRate and Kappa across all classes, i.e., means, standard derivations, $10th$ percentiles, and update them cell-wise in the 3D array.

Step 3. Trade-Off point
I compute the trade-off points by Eq.3.8 and visualize the trade-off points by Eq.3.9 based on three pairs of marginal distributions, as shown in Figure 3. The trade-off points of CLIP and EfficientNet on CIFAR1-100 and ImageNet respectively are shown in Table 3.2 and Table 3.3.

It can be noted that, **(1) CLIP model does not outperform the EfficientNet model**. Comparing the trade-off points in Tables 3.2 and 3.3, CLIP's generalization bound exceeds EfficientNet's by up to 0.16 on ImageNet, and its diversity bound is higher by up to 0.01. On CIFAR-100, CLIP's generalization bound is lower by up to 0.05, while its diversity bound is higher by up to 0.02. Although the CLIP's SSIM(lower bound) and

MODEL TYPE	CLIP	EFFICIENT NET
GENERALIZATION BOUND	0.364	0.206
DIVERSITY BOUND	0.087	0.075
SSIM(lower bound)	0.779	0.891
ZEROSHOT(upper bound)	0.175	0.106
MODEL SIZE(lower bound)	167M	23M

Table 3.2: TradeOff points on ImageNet

MODEL TYPE	CLIP	EFFICIENT NET
GENERALIZATION BOUND	0.852	0.902
DIVERSITY BOUND	0.164	0.139
SSIM(lower bound)	0.824	0.976
ZEROSHOT(upper bound)	0.228	0.166
MODEL SIZE(lower bound)	56M	43M

Table 3.3: TradeOff points on CIFAR-100

ZeroShot(upper bound) are better than EfficientNet's, EfficientNet's model size is much smaller than CLIP's.

Comparing the marginal distributions in Figure 3.15, the trends of CLIP and EfficientNet (including ErrorRate and Kappa) on SSIM and ZeroShot dimensions are similar (see the 1st and 2nd columns in Figure 3.15). However, the trends for CLIP are opposite to those for EfficientNet on the model size dimension (see the 3rd column). EfficientNet is a compact CNN architecture that uses a compound coefficient to scale models effectively, rather than randomly scaling width, depth, or resolution. Compared to the pretrained CLIP models, EfficientNet models are much smaller and more sensitive to changes in model size. Consequently, the CLIP model does not show an advantage against the EfficientNet model.

latex Copy code

A reasonable explanation is that the available pretrained CLIP models include both CNN and Transformer types. Here, I selected CNN-based pretrained CLIP models, but ViT-based CLIP models might perform better.

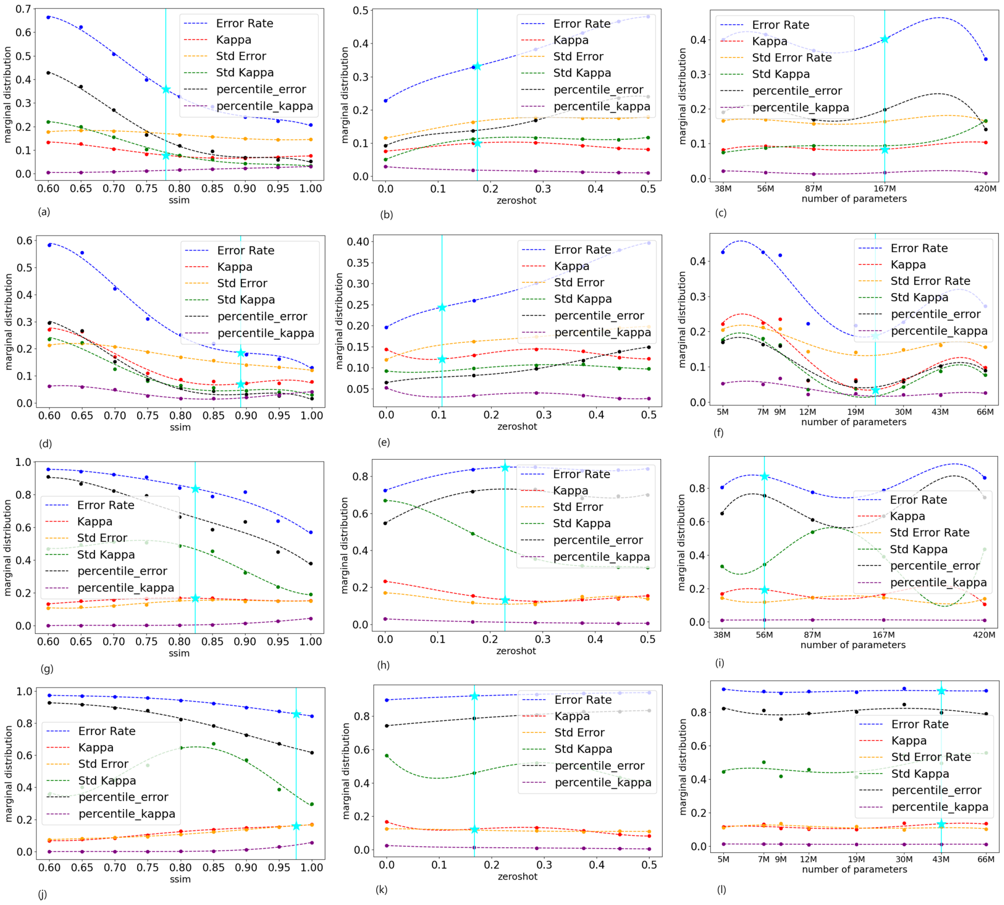


Figure 3.14: TradeOff points of two kinds models, CLIP and EfficientNet (denoted as " $\star$ "). The solid vertical lines indicate the selection of trade-off points on each marginals. (a)-(c) CLIP on ImageNet, (d)-(f) EfficientNet on ImageNet, (g)-(i) CLIP on CIFAR-100, (j)-(l) EfficientNet on CIFAR-100

(2) difference between datasets. It can be noted that the generalisation and diversity bounds on ImageNet are much less than on CIFAR-100 in Table 2 and 3. Moreover, it can be noted that STD Kappas on CIFAR-100 are obviously more than those on ImageNet in Figure 2. This indicates that the results on ImageNet are always better than on CIFAR-100 since big images can provide more data.

3.3.4 Consistency check with existing Generalisation Estimations

Dziugaite et al. (2020) and recent work (Sanae Lotfi (2023)) present 23 generalization measures, which I apply to all the pre-trained models listed in Table 3.1. my goal is to assess the consistency between existing theoretical estimations and actual measures, and to evaluate agreement/disagreement rates among the available theoretical approaches. For comparison, I focus on two slices of the 3D array rather than the entire array: one for data without robustness and another for data without zero-shot capacity. This allows me to obtain two distributions of error rates—one for robustness and model size dimensions, and the other for zero-shot and model size dimensions. Note that Kappa is not considered here, as the available complexity estimations focus on generalization error rates. I conduct the consistency check between theoretical estimations and actual measures using these two distributions.

The dimensions of robustness and zero-shot capacity are regarded as two independent factors. I compute two marginal probabilities of these two slices with respect to the dimension of $WeightNum$ (i.e., distributions with respect to $WeightNum$) as below,

$$\begin{cases} dtr_g(z \sim 2DSLICE(WeightNum)) = \\ \qquad \sum_{(y)\sim 2DSLICE(Robust)} dtr_g(y, z) \\ dtr_g(z \sim 2DSLICE(WeightNum)) = \\ \qquad \sum_{(x)\sim 2DSLICE(ZeroShot)} dtr_g(x, z) \end{cases} \tag{3.10}$$

Figure3.15(a)-(d) shows these marginals based on ImageNet and CIFAR-100 respectively. Then, I compute the empirical sign-error of generalization in terms of the resulting marginal probabilities Eq.3.10 as below,

$$\begin{aligned} SE_g = \tfrac{1}{2}\mathbb{E}_{(w,w')\sim\{WeightNum\}}\,[1 - sgn(dtr_g(w) \\ -dtr_g(w'))sgn(C(w) - C(w'))] \end{aligned} \tag{3.11}$$

where w and w' denote two different $WeightNum$s from the range of model size; $C(.)$ denotes the complexity measures computed using (Dziugaite et al. (2020), Sanae Lotfi (2023)). If the practical measures (dtr_g) and complexity measures (C) exhibit consistent changes, the sign-error (SE_g) approaches zero. Conversely, inconsistent changes lead to an SE_g approaching one. Consequently, an SE_g exceeding 0.5 indicates a potential mismatch between theoretical estimation and actual measures. Figure3.15(e)-(h) visualizes the distributions of sign-errors through scatter plots.

It can be noted that **most of generalisation bound estimations are not consistent with actual measures.**

Regarding the robustness dimension (SSIM), although Figure3.15e shows that 30% of SE_g error rates exceed 0.5, Figure3.15g indicates that all SE_g values are above 0.5. Furthermore, in both Figure3.15e and 3.15g, the SE_g values for the $10th$ percentile are all greater than 0.5, implying that the top-performing 10% of cases have an error rate exceeding 50%. This highlights a significant issue with the reliability of the estimation. For the ZeroShot dimension, Figure3.15f shows that 43% of SE_g error rates exceed 0.5, while Figure3.15h indicates that only 21% exceed 0.5. This suggests that the estimation performs better in the ZeroShot dimension compared to robustness. However, most of SE_g of $10th$ percentiles in Figure3.15f and 3.15h are still more than 0.3. The estimations' reliability is questionable.

3.4 Summary

This chapter, I propose an empirical generalization metric designed for deep networks. my approach involves a step-by-step illustration of the proposed generalization metric system, demonstrated through a series of tests on a singular model. I derive a quantifiable trade-off point that serves as a reliable indicator of the generalization capacity of the tested model.

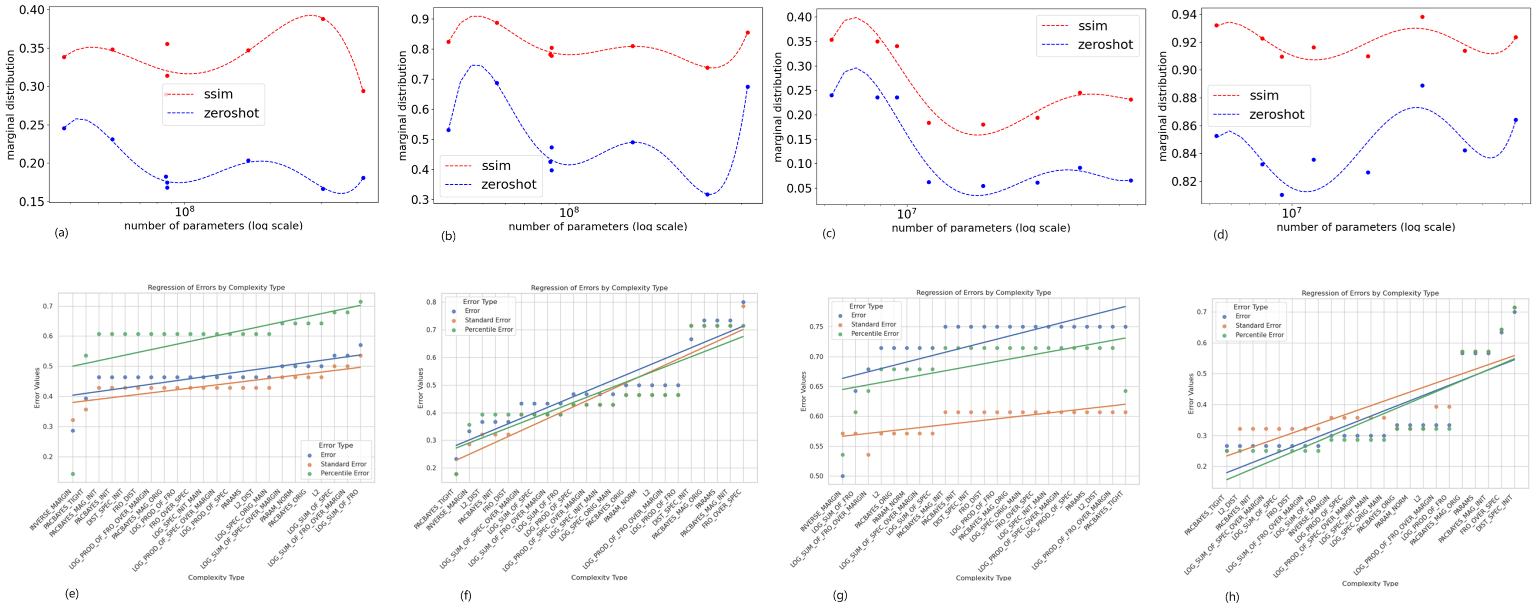


Figure 3.15: Upper row: Four marginal probabilities of two slices with respect to the dimension $WeightNum$: (a) CLIP (b) EfficientNet on ImageNet, (c) CLIP (d) EfficientNet on CIFAR-100. Bottom row: Scatter plots of the sign-errors: (e) related to SSIM on ImageNet, (f) related to ZeroShot on ImageNet, (g) related to SSIM on CIFAR-100, (h) related to ZeroShot on CIFAR-100.

Furthermore, I extend my examination to various SOTA model, benchmarking all models with respect to marginal probability (performance), randomness, robustness, model size, and zero-shot percentage. The versatility of the proposed metric is evident in its applicability for benchmarking a diverse range of deep networks.

In future, to enhance benchmarking, a broader range of architectures is required. I have initiated a public GitHub repository for deep network benchmarking and encourage contributions to expand the dataset and foster further theoretical and practical research. Furthermore, I will organise a comprehensive generalization benchmarking competition for deep networks. This future endeavor seeks to provide developers with a baseline platform to test new theories, thereby enhancing the understanding of why deep neural networks generalize. The benchmarking testbed will facilitate rigorous analyses, enabling developers to assess how well these theories align with the complexities observed in real-world models.

Chapter 4

Enhancing Generalization in Sketch-based Image Retrieval through Single Source Domain Adaptation

Sketch-based image retrieval (SBIR) is an active research area intersecting computer vision, multimedia, and machine learning. Traditionally, SBIR systems perform well when confined to a fixed dataset; however, their efficacy diminishes when exposed to new, unseen data categories, mirroring challenges in deep learning generalization. This issue is exacerbated when SBIR systems encounter test data from completely unfamiliar classes, revealing limitations in their ability to generalize beyond trained concepts. Such challenges highlight a critical need for robust domain adaptation strategies that enhance the model's generalization capabilities.

Current SBIR methodologies, while effective under controlled conditions, often falter due to two primary factors: the significant discrepancy between images and sketches and the varying abstraction levels introduced by different artist skills. This chapter expands upon the notion of domain adaptation in SBIR by proposing a novel framework that not only addresses the inherent disparities in SBIR but also enhances the system's adaptability across diverse datasets. This is achieved through the integration of canonical correlation analysis and advanced optimization techniques, enabling effective domain transfer and robust feature extraction.

Building directly upon the empirical generalization framework introduced in Chapter 3, this chapter implements a practical domain adaptation approach that applies these generalization principles to the specific context of SBIR. The model selection methodology established in Chapter 3 serves as the foundation for identifying optimal networks that can effectively bridge the domain gap between sketches and images, while the domain adaptation techniques presented here provide the mechanism for transferring knowledge between domains.

my main contributions are twofold. Firstly, I introduce a single-source domain adaptation algorithm for SBIR that facilitates effective transfer learning from one domain to another, enhancing the system's ability to handle new classes with minimal computational overhead. Secondly, I employ low-rank matrix decomposition and nonlinear approximation methods to significantly reduce computational complexity while maintaining high adaptation fidelity. This approach not only bridges the gap between theoretical advancements and practical applications in SBIR but also sets a foundation for future explorations into adaptive, generalizable retrieval systems.

4.1 Methodology

4.1.1 Overview of the Single Source Domain Adaptation Approach

This section presents my single-source domain adaptation methodology for SBIR, which directly implements the generalization principles established in Chapter 3. my approach addresses the challenge of transferring knowledge from data-rich source domains to target domains with limited labeled samples—a fundamental requirement for enhancing generalization in SBIR systems. The method integrates three key technical components: transfer learning through canonical correlation analysis, dictionary learning principles in the optimization process, and low-rank matrix decomposition for computational efficiency.

Figure 4.1 illustrates the architecture of my proposed approach. The workflow begins by identifying a target domain T with a structure similar to the new domain Y for which I have limited examples. I then employ canonical correlation analysis (CCA) to develop a model based on the relationship between the source and target domains. Finally, I apply this model to the new domain Y, using the few available labeled examples to fine-tune the adaptation.

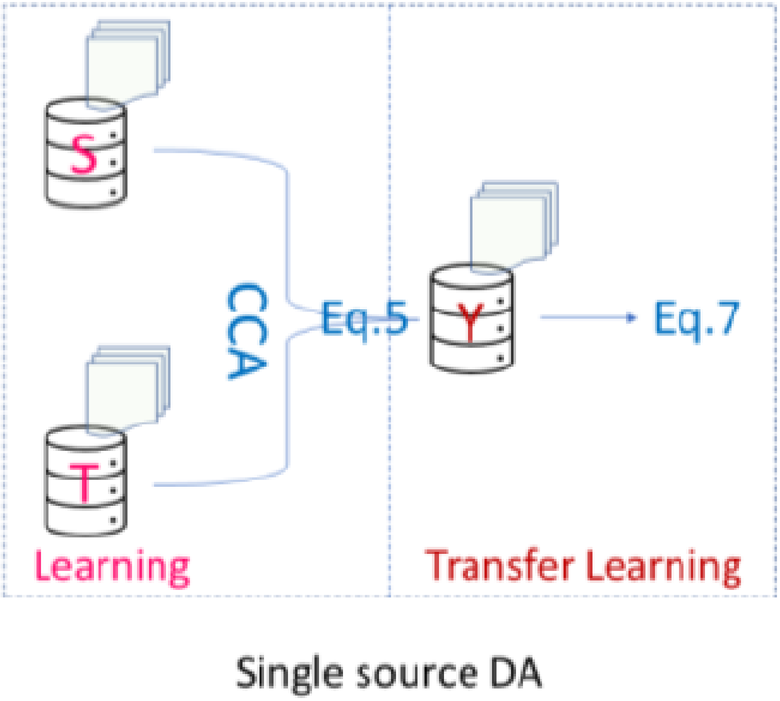


Figure 4.1: The structure of the proposed single source domain adaptation method. The process involves three key steps: (1) identifying a suitable target domain with structure similar to the new domain, (2) employing canonical correlation analysis to learn domain-invariant representations, and (3) transferring the model to the new domain with few-shot adaptation.

4.1.2 Transfer Learning via Canonical Correlation Analysis

Transfer learning, a core component of my approach, allows me to leverage knowledge gained in a source domain with abundant data to improve performance in a target domain with limited data. Unlike conventional transfer learning methods that typically involve fine-tuning entire networks, my approach identifies and transfers the essential statistical relationships between domains through canonical correlation analysis.

Consider a scenario with one source domain S and one target domain T, each containing k categories. For the i-th category, I can express the feature representations of examples in both domains as $S^i \in \mathbb{R}^{n_1 \times f}$ and $T^i \in \mathbb{R}^{n_2 \times f}$, $i = [1..k]$, respectively. Here, n_1 and n_2 represent the number of samples, and f denotes the feature dimension. I assume the features are unit-norm and $n_1 > n_2$, reflecting the common scenario where labeled samples in the target domain are limited.

The fundamental challenge in transfer learning for SBIR is to identify transformations that align the source and target domains while preserving the discriminative information necessary for accurate retrieval. Canonical correlation analysis provides a principled mathematical framework for this alignment by finding linear transformations that maximize the correlation between the two domains.

I apply CCA to each class of data as follows:

$$C_{SS}^{-\frac{1}{2}} C_{ST} C_{TT}^{-\frac{1}{2}} = LDR^T \in \mathbb{R}^{n_1 \times n_2} \tag{4.1}$$

where $C_{ss} = S^i S^{i^T}$, $C_{ST} = S^i T^{i^T}$, and $C_{TT} = T^i T^{i^T}$, $i \in [1...k]$ represent the covariance matrices. The matrices L and R contain the canonical vectors, while D is a diagonal matrix of canonical correlations that quantifies the strength of relationship between the domains.

The canonical correlation eigenspaces can be defined as:

$$\begin{cases} L_S = C_{ss}^{-\frac{1}{2}} L \in \mathbb{R}^{n_1 \times n_2} \\ R_T = C_{TT}^{-\frac{1}{2}} R \in \mathbb{R}^{n_2 \times n_2} \end{cases} \tag{4.2}$$

such that:

$$L_S^T C_{ST} R_T = D \in \mathbb{R}^{n_2 \times n_2} \tag{4.3}$$

Computing the projections of the sets S^i and T^i onto these eigenspaces yields:

$$\begin{aligned} P_s &= S^{i^T} L_S \in \mathbb{R}^{f \times n_2} \\ P_T &= T^{i^T} R_T \in \mathbb{R}^{f \times n_2} \end{aligned} \tag{4.4}$$

These projections, P_S and P_T, represent transformed feature spaces where the correlation between source and target domains is maximized. This transformation effectively bridges the domain gap between sketches and images by identifying the shared subspace where both domains are most aligned.

4.1.3 Dictionary Learning and Sparse Optimization

After obtaining the canonical correlation projections, I establish a mapping between the source and target eigenspaces:

$$P_S Q = P_T \tag{4.5}$$

where Q is a transformation matrix that can be initially estimated using the pseudoinverse of P_S:

$$Q = P_S^+ P_T \tag{4.6}$$

my approach incorporates principles from dictionary learning—a technique where signals are represented as sparse linear combinations of basis elements (atoms) from a learned dictionary. In my context, I view the transformation matrices as dictionaries that enable sparse representations of the domain adaptation mapping. This dictionary learning perspective offers several advantages:

1. Improved generalization by capturing only essential domain relationships 2. Robustness to noise and variations in sketching styles 3. Efficient representation of complex domain transformations

For a pair of highly correlated samples $\hat{S}$ and $\hat{T}$, I expect the correlation of their projections onto the eigenspace P_T to approach one:

$$\left\langle \hat{S}^T P_S Q, \hat{T}^T P_T \right\rangle \approx 1 \tag{4.7}$$

For the mean $\bar{S}^i \in \mathbb{R}^f$ of the set S^i in the source domain, I expect the correlations of the projections of $\bar{S}^i$ and the set T^i onto the eigenspace P_T to be:

$$\left((\bar{S}^i)^T P_S Q^i\right) \left(T^i P_T\right)^T \approx (1, ..., 1)_{1\times n_2} \tag{4.8}$$

To enhance numerical stability and address the non-convexity of the optimization problem, I introduce a new variable Ω and formulate the optimization as:

$$\min_{\Omega} \left\| \overrightarrow{1} - \left(T^i P_T\right) \Omega^i \left(\bar{S}^{i^T} P_S Q^i\right)^T \right\|, i = 1..k, \Omega^i \in \mathbb{R}^{n_2 \times n_2} \tag{4.9}$$

This allows me to estimate the mean of the set T^i in the target domain as:

$$\bar{T}^{i^T} = \bar{S}^{i^T} P_s Q^i \Omega^i P_T^+, i = 1..k \tag{4.10}$$

where P_T^+ is the pseudoinverse of P_T, and $\overrightarrow{1}$ denotes a vector of ones.

Equation 4.9 requires solving for a matrix Ω rather than a vector, presenting a non-convex optimization problem that may be susceptible to noise from computational processes or the inherent variability in sketches. To address this challenge, I introduce a sparsity constraint on Ω, formulated as $\min \|\Omega^i\|_0$, and employ the orthogonal matching pursuit (OMP) method (Pati et al. 1993) for solving this sparse optimization problem.

The OMP algorithm is particularly suitable for my application because it progressively identifies the dictionary atoms most correlated with the current residual without requiring preset sparsity parameters or error limits. This progressive selection process aligns with my goal of capturing only the most relevant domain relationships while discarding noise.

Through this sparse optimization procedure, I effectively establish a k-mean classifier based on the source and target domains S and T according to Equation 4.10.

4.1.4 Low-Rank Matrix Decomposition for Efficient Domain Adaptation

Low-rank matrix decomposition forms a critical component of my approach, addressing the computational challenges associated with high-dimensional feature spaces while ensuring robust transfer between domains. When working with large-scale retrieval systems, the transformation matrices can become prohibitively large, leading to computational inefficiency and potential overfitting.

Low-rank approximation addresses these challenges by:

1. **Dimension reduction**: Representing high-dimensional data in a lower-dimensional subspace 2. **Noise filtering**: Capturing only the most significant patterns in the data while discarding noise 3. **Computational efficiency**: Reducing the number of parameters that need to be optimized 4. **Improved generalization**: Limiting model complexity to avoid overfitting

my implementation employs the orthogonal matching pursuit (OMP) algorithm as a form of low-rank decomposition. By progressively selecting the most relevant components, OMP effectively produces a low-rank approximation of the transformation matrices. This approach is particularly valuable in SBIR applications, where the inherent variability in sketching styles introduces noise that can degrade retrieval performance.

4.1.5 Transfer to New Domains with Few-Shot Samples

To extend my approach to new domains with limited labeled examples (the few-shot scenario that is critical for practical SBIR applications), I consider a domain Y with the same k classes as the source domain S. Let the few-shot sample sets from Y be $Y^i \in \mathbb{R}^{n_2 \times f}, i = 1...k$. I transfer the classifier from Equation 4.10 to the new domain Y by updating Q^i as:

$$Q^i = Q^i \Omega^{i^T} \tag{4.11}$$

I then introduce a new variable Ω for Y^i into Equation 4.9 and solve:

$$\min_{\Omega} \left\| \overrightarrow{1} - \left(Y^i P_T\right) \Omega^i \left(\bar{S}^{i^T} P_S Q^i\right)^T \right\|, i = 1...k, \Omega^i \in \mathbb{R}^{n_2 \times n_2} \tag{4.12}$$

This leads to a similar estimation of the mean:

$$\bar{Y}^i = P_T \Omega^i Q^{i^T} P_S^T \bar{S}^i, i = 1..k, \bar{Y}^i \in \mathbb{R}^f \tag{4.13}$$

The classifier from Equation 4.10 is thereby transferred to the new domain Y, resulting in a new classifier (Equation 4.13) based on the source domain S and the new domain Y. Compared to Equation 4.10, Equation 4.13 offers the distinct advantage of bypassing the domain adaptation procedure from Equation 4.5, saving computational time. However, this efficiency may come at the cost of accuracy, as analyzed in Section 4.2.

4.1.6 Progressive Approximation for Convergence

my progressive approximation scheme draws inspiration from the progressive iteration approximation property (Lin et al. 2005). To compute the initial mapping between modalities (Q in Equation 4.5), I use the samples in S^i and T^i as constraints. I then update the constraint S^i with $\bar{S}^i$ and incorporate Q as a new constraint in Equation 4.9.

With sufficient constraints to update the mapping—for instance, by selecting arbitrary subsets of S^i as constraints and iteratively solving for a new Ω while absorbing the previous Ω into Q—the resulting mapping converges to the mapping between the means S^i and $\bar{S}^i$ in Equation 4.10.

This presents a non-convex optimization problem since the basis $P_S Q \Omega P_T^T$ in Equation 4.9 is positive semidefinite and likely rank-deficient. While Theorem 2.1 in (Lin et al. 2005) requires a nonsingular basis, I address this challenge by applying the low-rank decomposition technique OMP to minimize Equation 4.9, ensuring the mapping converges effectively.

The integration of transfer learning, dictionary learning, and low-rank matrix decomposition creates a comprehensive framework for domain adaptation in SBIR that directly implements the generalization principles established in Chapter 3. This methodology enables effective knowledge transfer from source to target domains while maintaining computational efficiency—a critical requirement for practical SBIR applications.

4.2 Experiment

This section first present the Implementation details including datasets and evaluation protocol. Then I compare my single source DA-SBIR method's performance with different other state-of-art algorithms in the normal few shot SBIR scenario. At last I evaluate the performance when I apply the method in Database-to-Database scenario.

4.2.1 Implementation details

Here, I provide the experimental details to evaluate the efficacy of the proposed approach for sketch-based image retrieval application. I start with the datasets used and evaluation details.

Datasets Used: In this work, I use three benchmark domain adaptation datasets.

The Sketchy Dataset (Sangkloy et al. 2016) is a large collection of sketch-photo pairs. The dataset consists of images from 125 different classes, with 100 photos each. Sketch are collected via crowd sourcing, which resulted in 75,471 sketches. This dataset also contains a fine-grained correspondence (aligned) between particular photos and sketches as well as various data augmentations for deep learning-based methods.In order to fit the task of large-scale SBIR, Liu et al. (2017) extended the dataset by adding 60,502 photos from Imagenet(Deng et al. 2009) yielding in total 73,002 images. I randomly pick 25 classes of sketches and images as the un-seen test set for the zero-shot SBIR, and the data from remaining 100 seen classes are used for training.

The TU-Berlin Dataset (Eitz et al. 2012) contains 250 categories with a total of 20,000 sketches extended by Zhang et al. (2016) with natural images corresponding to the sketch classes with a total size of 204,489. 30 classes of sketches and images are randomly chosen to respectively form the query set and the retrieval gallery. The remaining 220 classes are utilized for training. I follow Shen et al. (2018) and select classes with at least 400 images in the test set.

Feature Extraction: my feature extraction part is flexible and for the comparison I choose Doodle to search (Dey et al. 2019) method for single source domain adaptation on Sketchy and TU-Berlin dataset.

Evaluation protocol: The proposed evaluation uses the metrics used by Yelamarthi et al. (2018). Moreover, I also provide metrics on the whole dataset. Images labelled with the same category as that of the query sketch, are considered as relevant. Note that this evaluation does not consider visually similar drawings that can be considered correct by human users. For the existing datasets, I used random splits which contains same classes in both datasets. The mean Average Precision (mAP@all) is the main metric I use in the following experiment. First, I introduce the definition of precision in information retrieval scenarios. The precision is defined as the ratio of the retrieved images that are relevant to user's query over the retrieved documents. It can be represented as below:

$$\text{precision} \ = \frac{|\ \{\text{relavant images }\} \cap \{\text{ retrieved images }\}\ |}{|\ \{\text{ retrieved images }\}\ |} \tag{4.14}$$

By default, precision takes all the retrieved documents into account, but however, it can also be evaluated at a given number k of retrieved documents, commonly known as cut-off rank, where the model is only assessed by considering only its top-k-most queries. The measure is called precision at k or P@k. The Average Precision can be described as below:

$$AP@n = \frac{1}{GTP}\sum_{k=1}^{n} P@k \times rel@k \tag{4.15}$$

where GTP refers to the total number of ground truth positives, n refers to the total number of documents you are interested in, P@k refers to the precision@k and $rel@k$ is a relevance function. The relevance function is an indicator function which equals 1 if the document at rank k is relevant and equals to 0 otherwise. For each query, i, I can calculate a corresponding AP. The mAP is the mean of all the queries that the use made.

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{4.16}$$

mAP@all means n equals to the number of total images in the above equation. The second metric I utilize is classification rate. I classify the data in target domain after the domain adaptation and check the classification accuracy of each class.

4.2.2 Domain adaptation between images and sketches

The domain gap between image and sketch is the obstruction which strongly effect the retrieval result. I choose the image as the source data and sketch as the target. I build up a K-mean classifier according to Eq4.10 for SBIR applications. The key point is to assess the means of images and sketches in each class. I compare my methods, DA-SBIR algorithms, with the ZSIH (Shen et al. 2018),ZS-SBIR (Yelamarthi et al. 2018), SEM-PCYC (Dutta and Akata 2020) and DSN (Wang et al. 2021) methods on the two datasets: Sketchy and TU-Berlin. I employ the feature representation in the Doodle to search (Dey et al. 2019) I test the mean-average precisions (mAP@all) for all method and compare its result.

A concern is emerging regarding how the quantity of few-shot examples impacts retrieval performance. Figure 4.2 shows the few-shot SBIR performance of my model Eq.4.10 on the Sketchy and TU-Berlin databases respectively comparing with other methods. It can be noted that the results converge around 4 shot samples, and the DSN's result is close to ours on the Sketchy. I hope to point out that the DSN employs semantic information for retrieval. A pretrained Word2Vec (Mikolov et al. 2013) or GloVe (Pennington et al. 2014) network is required. However, my method Eq.4.10 still outperforms others though it does not involve semantic information. This can be noted in Table 4.1 as well, which illustrates the mAP@all and precision@100 with 20 few-shot samples. Moreover, the second issue is whether the feature representation influences the performance of algorithms. I take each method's feature representation as my method Eq.4.10's in turn and carry out 1-by-1 comparisons in Figure 4.3. It can be noted that (1) my method Eq.4.10 outperforms the others (see the dashed lines); (2) the feature representation heavily influences the methods' performance. This is not surprised since the performance of the classifiers (the neural networks such as ZSIH (Shen et al. 2018),ZS-SBIR (Yelamarthi et al. 2018), SEM-PCYC (Dutta and Akata 2020) and DSN (Wang et al. 2021) methods on the two datasets: Sketchy and TU-Berlin. I employ the feature representation in the Doodle to search (Dey et al. 2019)) always depends on the feature extraction and representation. Additionally, the feature representation of the Domain-Aware SE Network (Lu et al. 2021), which is applied to my method Eq.4.10 in Figure 4.3, does not involve semantic information. The feature representations of four existing methods contain semantic information, which are applied to my method Eq.4.10 respectively in Figure 4.3 (see the dashed lines). It can be noted that applying the feature representation of the Domain-Aware SE Network to my method Eq.4.10 achieves the results comparable with the state-of-the-art (see the dashed lines of using the DSN feature in Figure 4.3). There is much room for improvement of the feature representation. It is meaningful to separate the feature representation from the classifier design in SBIR applications.

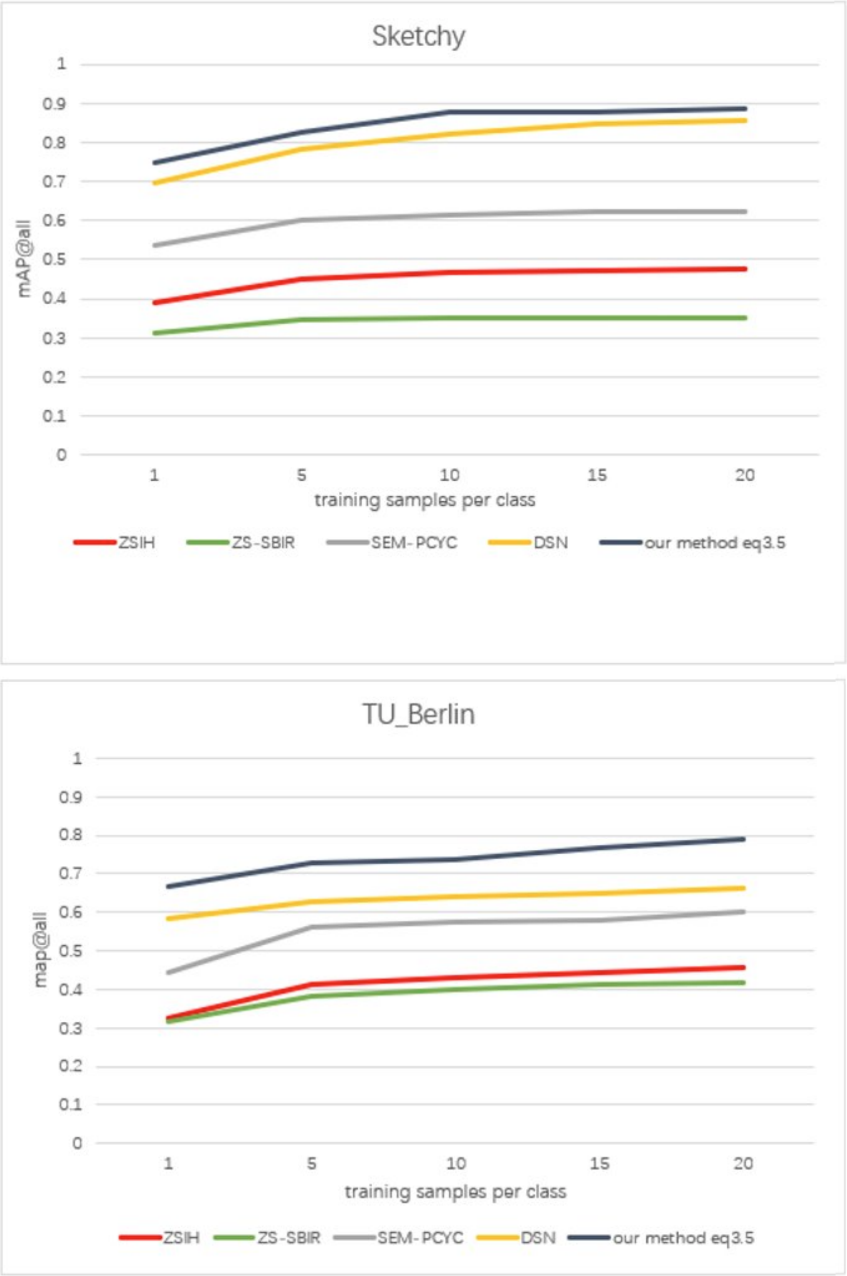


Figure 4.2: The illustration of the influence of the few-shot sample number in the target domain to retrieval performance.

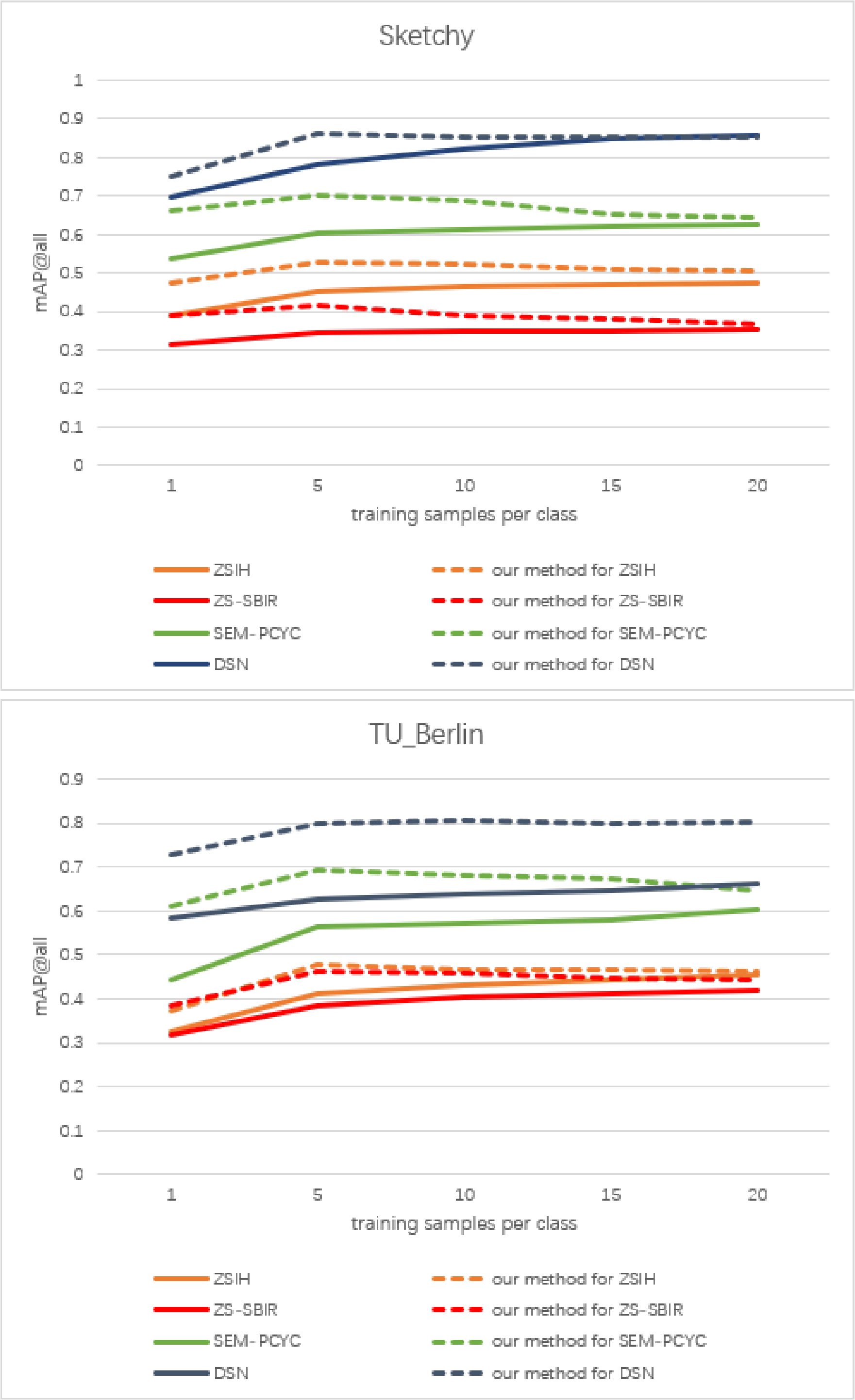


Figure 4.3: 1-by-1 performance comparison of each method with my method (Eq.4.10). (The dashed lines indicate my method's performance.

Table 4.1: COMPARISON OF Methods' Performance when the number of few-shots is of 20.

	Sketchy		TU-Berlin	
	mAP@all	precision@100	mAP@all	precision@100
ZSIH	0.4527	0.5728	0.4523	0.5837
ZS-SBIR	0.3587	0.4782	0.4122	0.5462
SEM-PCYC	0.6063	0.7387	0.6005	0.7283
DSN	0.8574	0.9764	0.6631	0.7927
my Method Eq.4.10	0.8863	0.9813	0.7882	0.8827

Table 4.2: The worst/best mAP@all of my method and their distances between the estimated and the ground truth means on the TU-Berlin database. (Note that the few-shot number is 4)

mAP@all / distance of means	ZSIH	ZS-SBIR	SEM-PCYC	DSN	my method
Worst class	0.2759	0.1275	0.4736	0.5837	0.1364
Best class	0.8758	0.6927	0.9647	0.9826	0.9321
Average	0.4073	0.3964	0.5572	0.6129	0.6037
Distance between Estimated and Ground Truth (worst class)	0.9473	0.9972	0.8863	0.8037	0.9382
Distance between Estimated and Ground Truth (best class)	0.5729	0.7746	0.3974	0.3472	0.3863

The third issue refers to increasing the number of few-shots does not improve the retrieval performance significantly. To address this issue, I take 4 labeled sketches from the target by default to satisfy the few-shot settings based on the TU-Berlin dataset and further check all the mAP@all of the classes. Table4.2 shows the best and worst performance using my method Eq.4.10. Moreover, I also plot the sample distributions of the best and worst classes by the t-SNE (Van der Maaten and Hinton 2008) in Figure 4.4. It can be noted that the sample distribution of the worst class is very dispersal and the few-shot samples are likely from "outliers". Usually few-shot samples are less than 5% of total samples in the target. There is no sampling approach to guarantee that few-shot samples are good representative of a statistical population. This also answers the above-mentioned question, that is, for few-shot scenarios, the limited number of few-shot samples in the target are likely from outliers and result in incorrect classification.

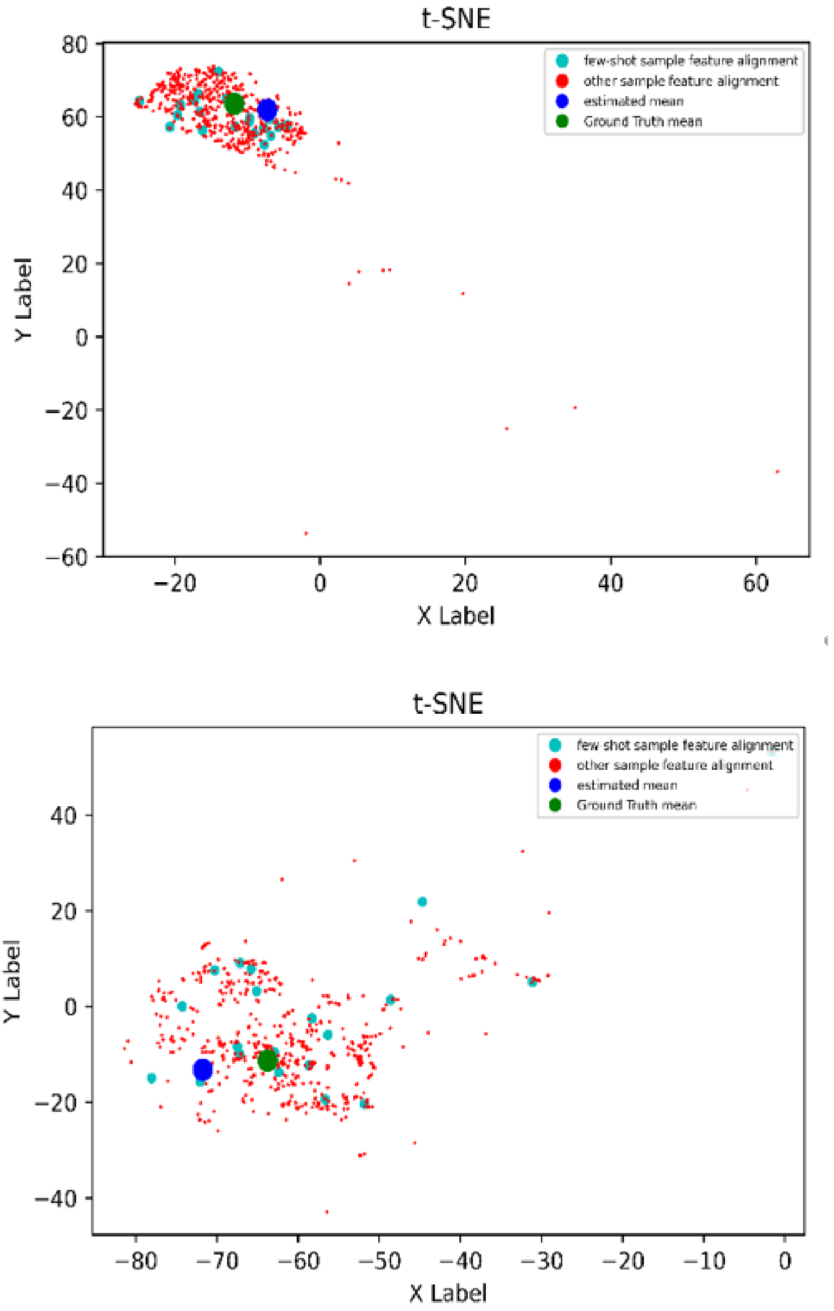


Figure 4.4: Illustration of the best class (top) and the worst class (bottom) sample distribution and the estimations of the means

Qualitative Results. Next, I analyze the retrieval performance of my proposed model qualitatively in Figure 4.5. Some notable examples are as follows. Sketch query of tank retrieves some examples of motorcycle probably because both of them have wheels in common. For having visual and semantic similarity, sketching guitar retrieves some violins. Querying castle, retrieves images having large portion of sky, because the images of its semantically similar classes, such as, skyscraper, church, are mostly captured with sky in background. In general, I observe that the wrongly retrieved candidates mostly have a closer visual and semantic relevance with the queried ones. This effect is more prominent in TU-Berlin dataset, which may be due to the inter-class similarity of sketches between different classes. Therefore, for TU-Berlin dataset, it is challenging to generalize the unseen classes from the learned representation of seen classes.

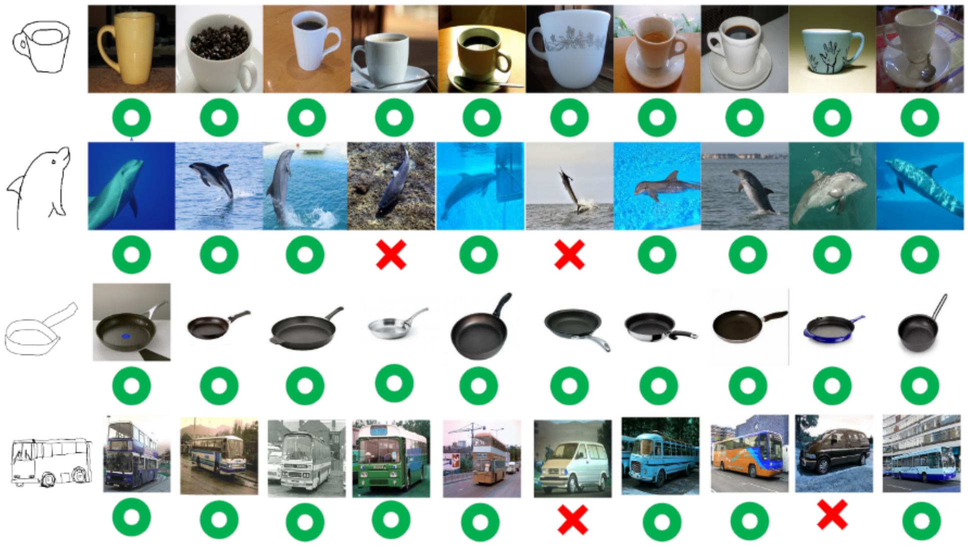


Figure 4.5: Top 10 image retrieval examples given a query sketch. All the examples correspond to a few-shot setting(20-shot). First two rows provides a retrieval result from Sketchy Dataset and last two rows shows the result of TU-Berlin Dataset. Note that in some retrieval cases, for instance, Dolphin is confused with fish images which can be true even for humans. Green circle and Red Cross stands for correct and incorrect retrievals.

4.2.3 Domain Adaptation between databases

In practice, new sketch-photo datasets continually become available for various SBIR applications. Transferring classifiers trained on well-established datasets to these new datasets is highly beneficial. However, these fresh datasets are typically characterized by their limited size, offering only a few labeled sketch-photo pairs per category, thus presenting a data scarcity issue. This scenario is often referred to as the few-shot problem. In my approach, I create a source/target domain configuration using the "Sketchy dataset" as the source and the "TU-Berlin dataset" as the new target domain. To tackle this challenge, I employ a K-mean classifier based on Eq.4.10 and adapt it to the new domain using only a small number of labeled samples from the new domain, emphasizing the domain adaptation problem. In contrast, I evaluate other existing methods exclusively on the "TU-Berlin dataset," where both the source and target samples are drawn from the same dataset. This is because these existing methods do not utilize the source/target learning approach and do not require the introduction of a new domain. Typically, data from the same dataset exhibits better consistency compared to data from different datasets. As a result, the domain adaptation problem may be trivial when data is from the same dataset. The workflow of the Domain Adaptation between databases is shown in Figure 4.6

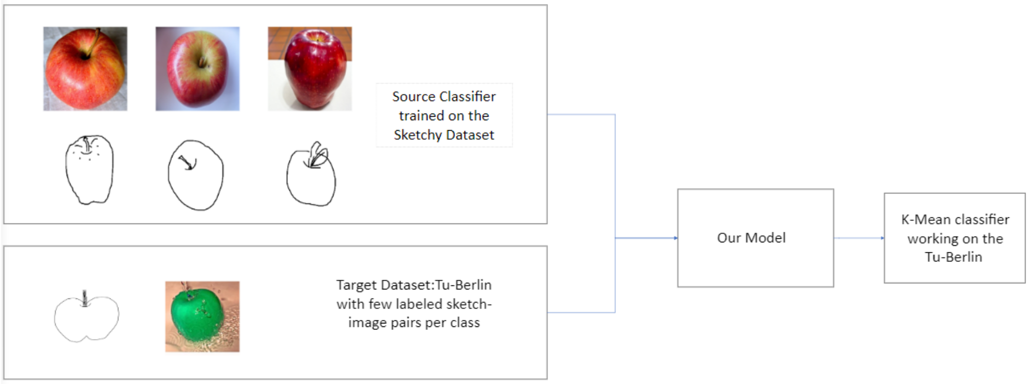

Source Classifier trained on the Sketchy Dataset
Target Dataset:Tu-Berlin with few labeled sketch-image pairs per class
Our Model
K-Mean classifier working on the Tu-Berlin

Figure 4.6: The workflow of the Domain Adaptation between databases

Firstly, I check the influence of the number of few shots to retrieval results in Figure 4.7. It implies again that increasing few-shot samples does not benefit retrieval performance. Thus, I choose the few-shot number as 4 by default. Moreover, it can be noted that although my method works across different datasets, the performance is comparable with the other methods working on a single dataset.

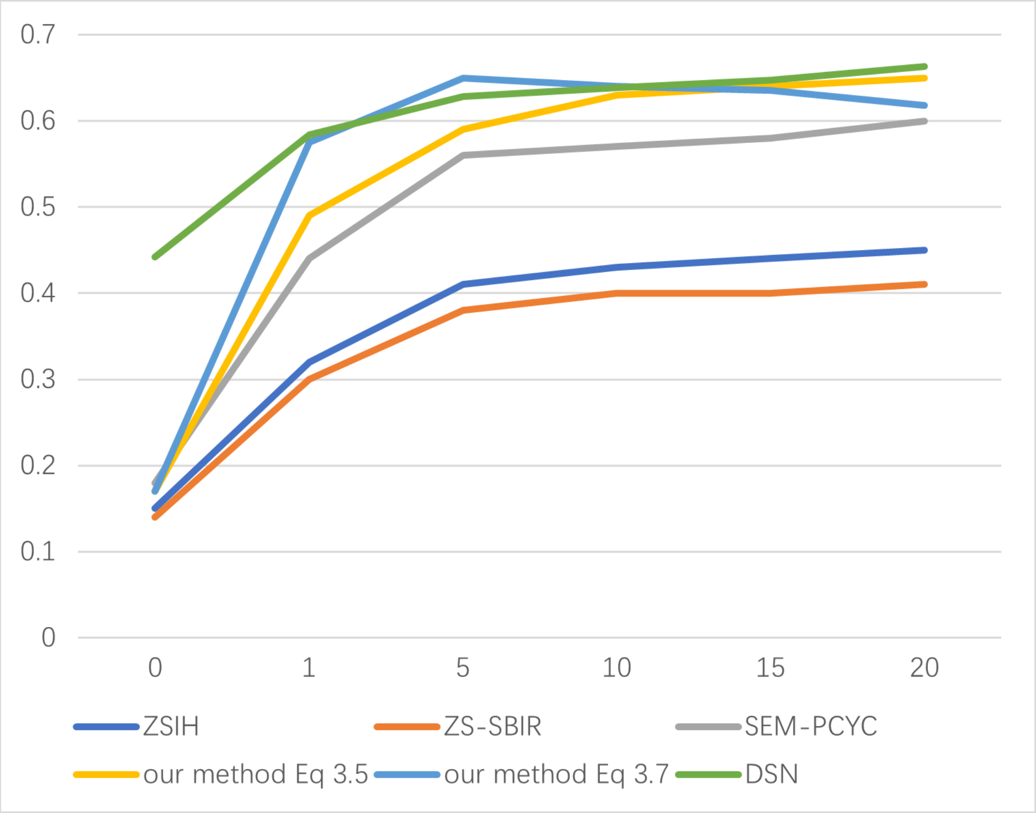


Figure 4.7: The influence of the number of few-shots to retrieval results.

Table 4.3 further shows the performance using 20 few-shot samples. It can be noted that my method still achieves competitive results compared with the state-of-the-art though the performance decreases slightly in the scenario of crossing datasets. Moreover, the classifier of Eq.4.13 saves computational time but does not cause a noticeable decrease in performance.

Table 4.3: PERFORMANCE OF CROSSING DATABASES

Source/Target:TU-Berlin - single dataset	mAP @ all	precision@100
ZSIH	0.4523	0.6037
ZS-SBIR	0.4122	0.5582
SEM-PCYC	0.6005	0.7283
DSN	0.6631	0.7927
my method (Eq.4.10)	0.6553	0.7818
my Method (Eq.4.13) - crossing datasets (Source:Sketchy vs. Target:TU-Berlin)	0.5867	0.7189

Secondly, I still find out the best and worst performance by my method of Eq.4.13 according to the distances between the mean's estimations and the ground truth means within all classes as shown in Table 4.4. I further visualize the sample distributions of two best classes and worst class by the t-SNE in Figure 4.8. Compared with Figure 4.4, it shows again that the limited number of few-shot samples in the target are likely from the outliers and results in incorrect classification. The selection of few-shot samples from the target indeed affects retrieval results. To further improve retrieval performance, it is most likely to involve text-base side information such as semantic space (Dutta and Akata 2020) (Wang et al. 2021).

Table 4.4: PERFORMANCE OF my METHOD (Eq.7) ACCORDING TO THE MEAN ESTIMATIONS. (THE FEW-SHOT NUMBER IS 4)

	distance(sketch)	accuracy@N (sketch)	distance(image)	accuracy (s (image)
Best class 1	1.23391	96.25%	0.831107	79.34%
Best class 2	1.0995	96.29%	0.5355	95.57%
Worst class	1.18755	5%	0660308	34.52%

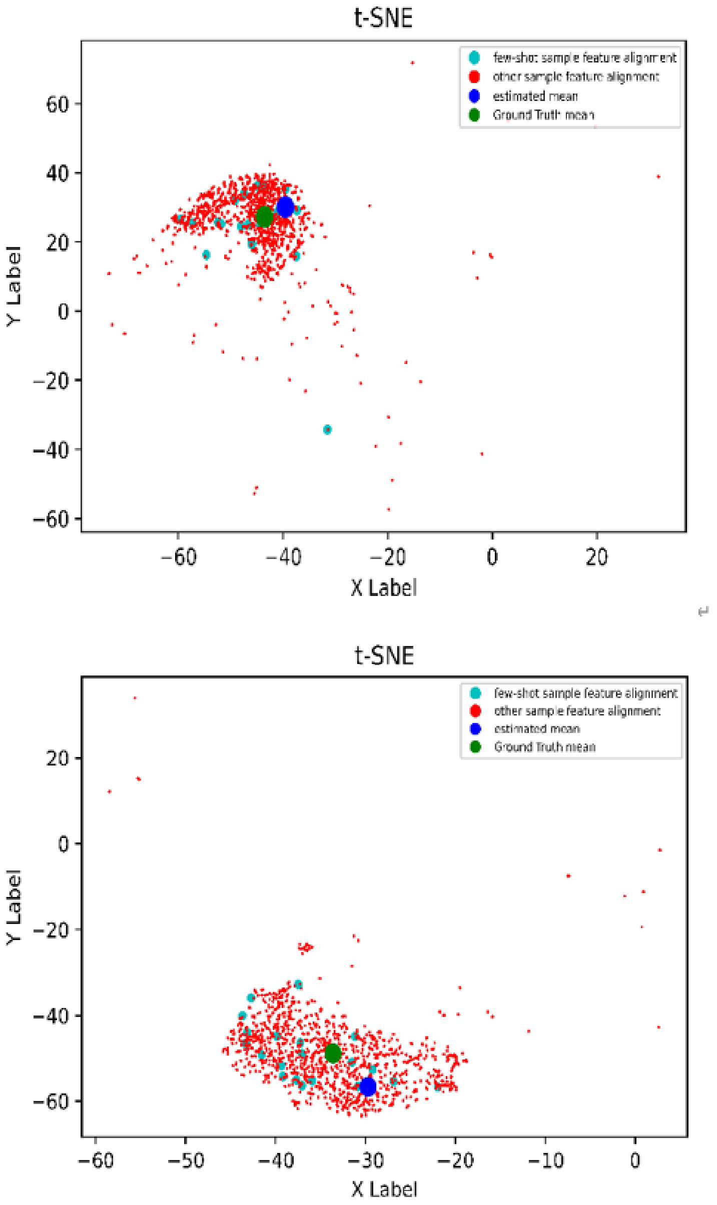


Figure 4.8: Illustration of the sample (images) distribution of the best class 2 (top) and worst class (bottom)

4.3 Summary

This chapter presents my SBIR algorithms for single source domain adaptation scenario. The main merits include, in few-shot scenarios deal with the cases of image-to-sketch domain adaptation and dataset-to-dataset domain adaptation. This is indeed to transfer learning models from one source domain to a target. Compared with the state-of-the-art methods, my algorithms do not use text based semantic information except sketches, but experiments on the Sketchy and TU-Berlin benchmark databases demonstrate that my algorithms achieve competitive results. This does not only show the effectiveness of my algorithms and also illustrate a promising perspective, that is, combination of deep networks with traditional machine learning techniques can bring about compelling performance.

Chapter 5

Multi-Source Domain Adaptation for Robust SBIR Systems

Building upon the single-source domain adaptation approach presented in Chapter 4, this chapter extends my generalization framework to the more complex yet practically valuable scenario of multi-source domain adaptation (MSDA). This extension directly implements the generalization principles established in Chapter 3 while addressing the limitations of the single-source approach in Chapter 4, creating a comprehensive solution for enhancing SBIR performance across diverse domains.

5.1 Advancing Generalization through Multi-Source Integration

Sketch-based Image Retrieval (SBIR) faces significant challenges in real-world applications due to the inherent differences in feature distributions and patterns between training (source domains) and testing (target domains) datasets. This discrepancy, known as the domain gap, is particularly pronounced in SBIR where sketches and photos can vary greatly not only in style but also in representational detail such as color and texture. While the single-source approach in Chapter 4 demonstrated promising results, it remains limited when the target domain exhibits characteristics that differ substantially from the single source domain.

my empirical generalization framework from Chapter 3 identified that both model accuracy and data diversity are critical factors in generalization performance. The multi-source domain adaptation (MSDA) methodology presented in this chapter directly addresses the data diversity component by leveraging multiple source domains to create a more robust representation space, thereby improving the model's ability to generalize across different visual representations. This approach represents a natural progression in my exploration of generalization enhancement for SBIR systems.

Traditional methods in SBIR often suffer from a lack of robustness when confronted with new classes or variations within the data. These methods typically employ discriminative modeling to align sketches with photographs by reducing intra-class variance and maximizing inter-class differences. However, such approaches can falter under the practical constraints of limited labeled data in the target domain. By contrast, my proposed MSDA framework utilizes multiple diverse source domains to enrich the model's ability to generalize across different visual representations, directly implementing the data diversity principles established in my generalization framework.

my MSDA approach not only addresses the challenge of integrating diverse data sources but also ensures that the retrieval system can effectively utilize unlabeled or sparsely labeled data from target domains. I extend the canonical correlation analysis approach from Chapter 4 and integrate it with online dictionary learning to minimize computational complexity while maximizing domain adaptation effectiveness. The ultimate goal is to develop a retrieval system that maintains high performance across various data distributions by preserving domain-invariant features and effectively mitigating the domain gap, thus achieving superior generalization as measured by my empirical metric system.

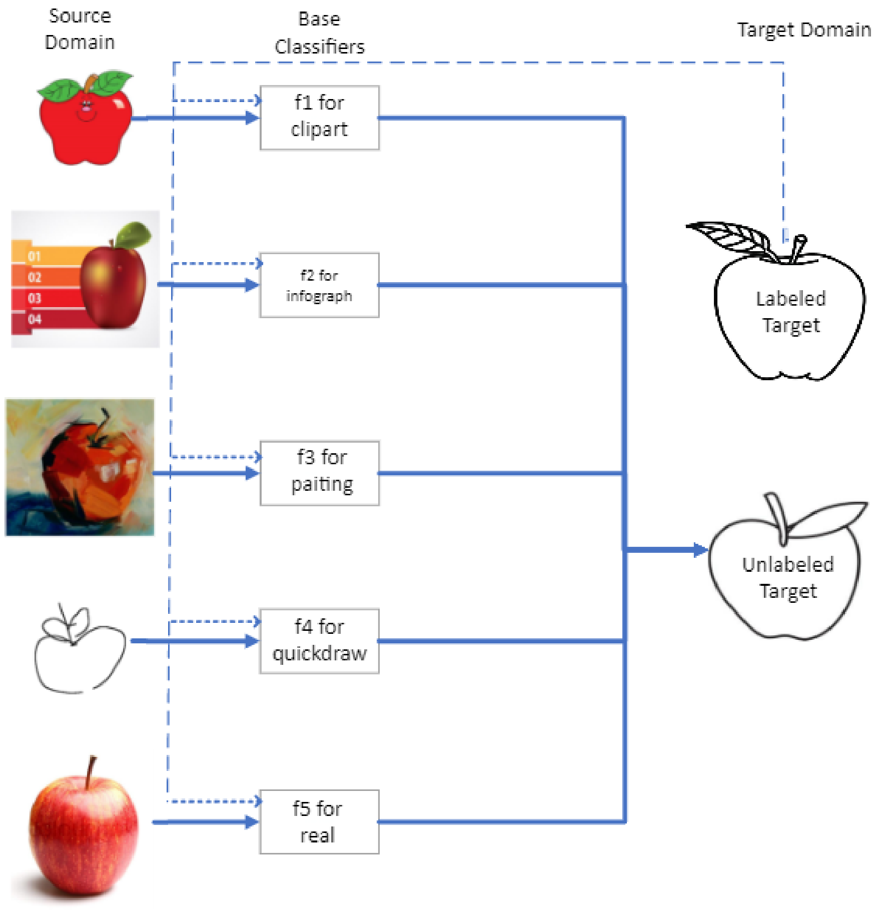


Figure 5.1: The structure of the multi-source domain adaptation approach for SBIR. The method extends the single-source framework from Chapter 4 by integrating multiple source domains through a shared canonical correlation space, enabling more robust generalization to target domains with limited labeled samples.

5.2 Methodological Advancement and Contributions

This chapter methodologically advances my generalization framework through several key contributions:

Introduction of a Robust MSDA Framework for SBIR: I propose a novel multi-source domain adaptation approach that significantly enhances the generalization capability of SBIR systems by effectively leveraging multiple source domains. This directly extends the single-source approach from Chapter 4 to address more complex real-world scenarios where diverse training data is available.

Computational Efficiency: Building upon the low-rank matrix decomposition techniques introduced in Chapter 4, I incorporate online dictionary learning to handle the increased complexity of multi-source scenarios. These techniques not only improve adaptation accuracy but also maintain computational efficiency, making my framework suitable for large-scale applications. This addresses the practical challenges identified in my generalization framework regarding model complexity and performance trade-offs.

Practical Applicability: By addressing both theoretical and practical aspects of domain adaptation, my approach bridges the gap between academic research and real-world SBIR applications, providing a robust framework that can be readily implemented in various settings. This aligns with my overall goal of enhancing generalization in practical SBIR systems, as established in Chapter 3.

The multi-source approach presented in this chapter represents a significant advancement over the single-source method introduced in Chapter 4. While both approaches utilize canonical correlation analysis and low-rank matrix decomposition, the MSDA framework introduces additional techniques to handle the increased complexity of integrating multiple source domains:

1. The single-source approach establishes domain transfer between one source and one target domain, whereas the MSDA framework must effectively combine information from multiple source domains while filtering out domain-specific noise.

2. The multi-source approach requires more sophisticated optimization techniques to balance the contributions of different source domains, necessitating the integration of online dictionary learning with my canonical correlation framework.

3. The computational complexity increases substantially with multiple source domains, requiring additional efficiency measures beyond those employed in the single-source approach.

my proposed multi-source DA-SBIR algorithms adopt low-rank matrix decomposition technology as a linear approximation method to keep computational complexity manageable. While it is feasible to implement these algorithms using nonlinear approximation approaches such as deep networks to potentially improve performance, my focus on linear methods aligns with the efficiency requirements identified in my generalization framework from Chapter 3.

The following sections detail my multi-source domain adaptation methodology, experimental validation, and analysis of results, demonstrating how this approach implements and extends my generalization framework to enhance SBIR performance across diverse domains.

5.3 Methodology

The diagrams of the suggested methods are shown in Figure 5.2 When dealing with a fresh field Y, my initial step involves identifying a target area T that shares a similar or identical structure with Y. Subsequently, I employ the source/target learning method with canonical correlation analysis (CCA) to build a model (eq4.10).

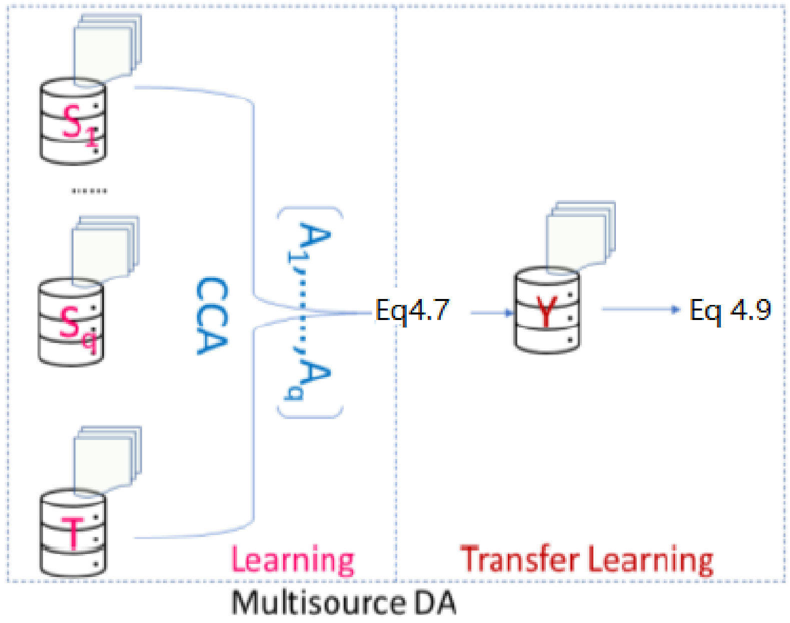


Figure 5.2: The structure of the proposed multi-source domain adaptation method

Consider the single source to a target scenario. From the chapter 4 I can conclude that to estimate the mean of the set T^i in the target domain as,

$$\bar{T}^{i^T} = \bar{S}^{i^T} P_s Q^i \Omega^i P_T^+, i = 1..k \tag{5.1}$$

In a situation where multiple source datasets and one target dataset were used, the goal was to determine the average values of the classes in the target dataset, based on information from the multiple source datasets. Initially, it made sense to use a specific equation Eq.5.1 separately on each of the source datasets and also on the target dataset. Then, the correlations can be expressed as,

$$\left(T^i P_T^{i,j}\right) \Omega^{i,j} \left(Q^{i,j^T} P_S^{i,j} \bar{S}^{i,j}\right), i = 1..k, j = 1..q \tag{5.2}$$

where T^i denotes the i-th class dataset in the target domain. Let $A^{i,j} = P_T^{i,j} \Omega^{i,j} Q^{i,j^T} P_S^{i,j^T} \in R^{f \times f}$. The correlations can be rewritten as,

$$T^i A^{i,j} S^{i,j} \tag{5.3}$$

Similar to Eq.4.9, I optimize a new variable α through,

$$\min_{\alpha} \left\| \overrightarrow{1} - T^i \left(A^{i,1}, \ldots, A^{i,q} \right) \alpha_i \begin{pmatrix} \bar{S}^{i,1} \\ \ldots \\ \bar{S}^{i,q} \end{pmatrix} \right\|_F^2, i = 1..k, \alpha_i \in R^{qf \times qf} \tag{5.4}$$

to estimate the mean of the i-th class in the target domain as,

$$\bar{T}^i = \left(A^{i,1}, \ldots, A^{i,q} \right) \alpha_i \begin{pmatrix} \bar{S}^{i,1} \\ \ldots \\ \bar{S}^{i,q} \end{pmatrix}, i = 1..k \tag{5.5}$$

The advantage is that the Eq.5.5 relies on all the source data in case of biased estimations. However, in a more realistic setting, Eq.5.4 is not feasible since the total size of source datasets may be too big to fit all the data into memory. Obviously, increasing the source number q results in a huge square matrix *alpha* in Eq.5.4 and memory overflowing quickly. Alternatively, I adopt the online dictionary learning (ODL) technology to deal with a large number of source datasets and lower the computational complexity. I firstly apply the ODL to,

$$\min_{D,\alpha} \sum_{j=1}^{q} \left\| A^{i,j} - D_i \alpha_i \right\|^2 + \rho \left\| \alpha_i \right\|_1, i = 1..k \tag{5.6}$$

where $D_i \in R^{f \times f}, \alpha_i \in R^{f \times f}$,to generate the dictionary D for each class. Theoretically, ODL can deal with any number of source domain datasets, i.e. q may be very big. However, it adopts traditional iterative batch procedures, which is prone to biased estimations. To tackle this deficiency, I still have to generate some linear combinations of the current and previous $A^{i,q}$ as additional source data and update D through Eq.5.6 for unbiased estimations. This can effectively reduce biases. Then, I apply sparse coding technology to,

$$\min_{\beta} \left\| \overrightarrow{1} - T^i D_i \left(\bar{S}^{i,1}, \ldots, \bar{S}^{i,q} \right) \beta_i \right\|, i = 1..k, \beta_i \in R^q \tag{5.7}$$

to determine the sparse weights *beta* to combine the class means from different source domain datasets together. The mean of the i-th class in the target domain is estimated by,

$$\bar{T}^i = D_i \left(\bar{S}^{i,1}, \ldots, \bar{S}^{i,q}\right) \beta_i, i = 1..k \tag{5.8}$$

Usually, when q is very big, the sparsity makes sense. So far, I set up the multisource classifier of Eq.5.8 based on q source domain data and one target domain data. Furthermore, like Eq.4.12, I can transfer the classifier of Eq.5.8 to the new domain Y with the same k classes as the T. Using the same skill, i.e.,

$$\left(\bar{S}^{i,1}, \ldots, \bar{S}^{i,q}\right) = \left(\bar{S}^{i,1}, \ldots, \bar{S}^{i,q}\right) \beta_i, \quad \beta_i \in R^q \tag{5.9}$$

I introduce a new variable β for each class Y^l and solve it by minimizing,

$$\min_{\beta} \left\| \overrightarrow{1} - Y^i D_i \left(\bar{S}^{i,1}, \ldots, \bar{S}^{i,q}\right) \beta_i \right\|, i = 1..k, \beta_i \in R^q \tag{5.10}$$

which yields a new classifier for the domain Y,

$$\bar{Y}^i = D_i \left(\bar{S}^{i,1}, \ldots, \bar{S}^{i,q}\right) \beta_i, i =| 1..k \tag{5.11}$$

5.4 Experiment

5.4.1 Implementation details

Here, I provide the experimental details to evaluate the efficacy of the proposed approach for sketch-based image retrieval application. I start with the datasets used and evaluation details.

Datasets Used: In this work, I use three benchmark DA datasets.

DomainNet (Peng et al. 2019) is a recently released large scale DA dataset with 6 different domains and a total of 345 classes with over 0.6 million images. Due to the prevalence of noise, experiments are conducted on the partial dataset consisting of 6 domains and 121 classes. This dataset I choose as the auxiliary dataset for multi-domain adaptation to help improve the retrieval result.

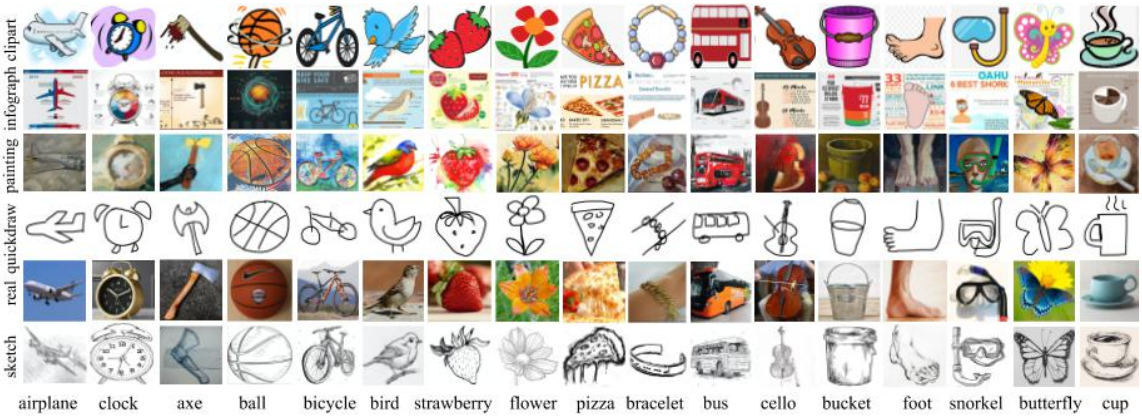


Figure 5.3: The Domainnet Dataset (Peng et al. 2019)

Feature Extraction: my feature extraction part is flexible and for the comparison I choose CUMIX method (Mancini et al. 2020) for multi-source domain adaptation on DomainNet dataset.

Evaluation protocol:I randomly pick training and test data from the above-mentioned datasets, containing same classes in both datasets, and employ the metrics of the mean Average Precision (mAP@all), precision considering top 100 (precision@100) and the classification accuracy as below,

$$\text{Accuracy@all } = \frac{N_{\text{relevant}}}{N}(\%) \tag{5.12}$$

5.4.2 Multisource Domain Adaptation experiment on DomainNet Dataset

I apply my method of Eq.5.8 to the DomainNet dataset for multisource domain adaptation applications. The DomainNet dataset has 6 different domains, including Clipart, Infograph, Painting, Quickdraw, Real, Sketch. In my experiments, I select the Sketch domain from the DomainNet dataset as my target domain, while the remaining five domains serve as the multi-source domains. Firstly, I examine the impact of the few-shot number on retrieval performance as depicted in Figure5.4. I notice a consistent peak performance with 3-4 few-shots, and encounter the same phenomenon where an increase in few-shot samples does not markedly enhance retrieval performance. Therefore, I consistently opt for a default few-shot number of 4 in my experiments. Secondly, I illustrate the retrieval results of my method (Eq.5.8) using different numbers of multisource domains from the DomainNet as shown in Table 5.1. I observe that as the number of source domains increases, the enhancement in retrieval performance tends to plateau at around 4 source domains. This suggests that the accumulation of noise from various domains could significantly hinder retrieval performance.

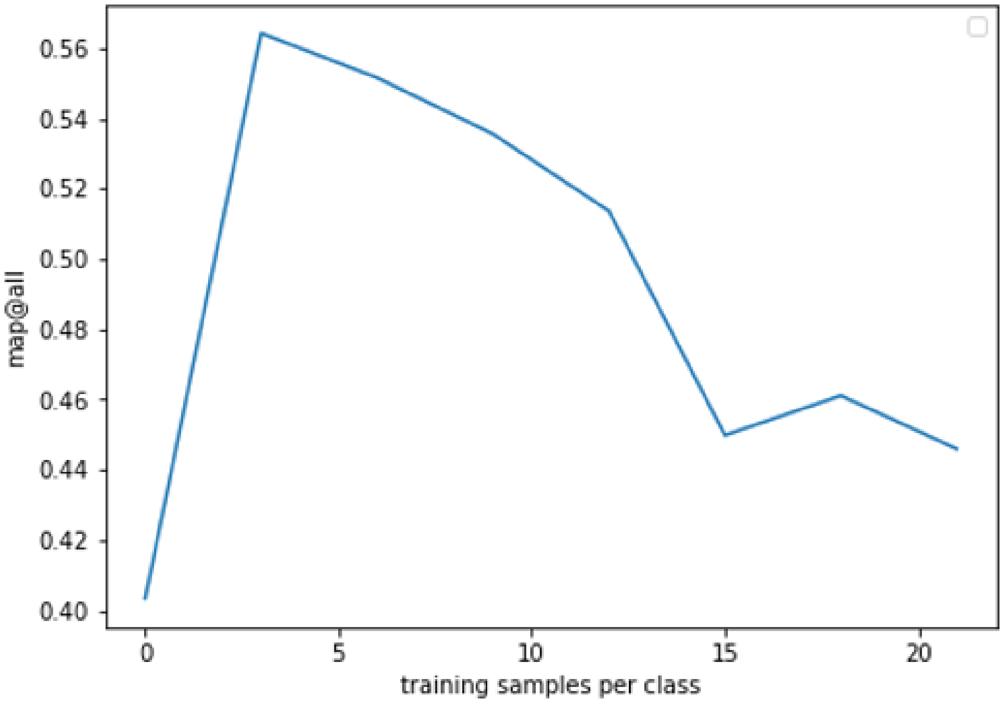


Figure 5.4: Influence of the few-shot number to retrieval results.

Table 5.1: Comparison of mAP@all when the number of few-shots is of 20.

Number of domains	1	2	3	4	5
mAP@all	0.2611	0.5003	0.5031	0.5772	0.4792
Accuracy@all	27.77%	54.70%	54.21%	59.49%	53.14%

I further compare my algorithm of Eq.5.8 with two existing multisource methods: DCTN (Xu et al. 2018) and M3SDA (Peng et al. 2019) on the DomainNet dataset and summarize their classification accuracies in Table 4.2. It is clear that multisource domains can evidently improve the classification accuracy compared with single source scenarios. The experiment of DCTN (Xu et al. 2018) and M3SDA (Peng et al. 2019) are conducted under unsupervised condition and cannot run under few shot settings. Thus, I compare my few-shot method with their original settings. my method of Eq.5.8 outperforms the other existing methods and avoids the negative transfer in the quickdraw domain.

Table 5.2: Comparison of classification accuracies before and after deleting the classes with the worst variances

Target domain	Clipart	infograph	painting	quickdraw	real	sketch	Average
DCTN	48.6%	23.5%	48.8%	7.2%	53.5%	47.3%	38.2%
M^{3}SDA	57.2%	24.2%	51.6%	5.2%	61.6%	49.6%	41.6%4
my method	54.4%	28.0%	41.4%	47.5%	53.9%	53.0%	46.4%4

Thirdly, a straightforward idea of further improving classification accuracy is to delete some seriously dispersal classes from the available multisource domains. However, the improvement is limited. These worst classes can be found out in terms of the variance of every class. For each class, I only delete the worst one within five source domains, so that each class has four sets of class data. For every class, Table 5.3 shows the classification accuracies and distances between the estimated and ground truth means before and after deleting the worst classes. The worst classification result before deleting is of the class 0, whose variances range 0.7546 to 0.9475. Deleting the class of 0.9475, the classification accuracy increases from 0% to 50.68%. Whereas the best classification result before deleting is of the class 38, whose variance ranges from 0.6680 to 1.100. Deleting the class of 1.100, the classification accuracy decreases from 100% to 86.92%. Although the overall performance is somewhat improved through deleting worst classes (see average), it can still be noted that the classification accuracies of some classes are decreasing. Moreover, it can also be noted that the classification accuracy of some classes remains very low throughout, e.g. the classes 1,15,18 and 31. This is because the sample distribution of these classes is too dispersal and the few-shot samples may be the outlies from the target. Thus, selecting few-shot samples from the target becomes very sensitive.

Table 5.3: Comparison of classification accuracies.(Note that the few-shot number is 4)

class index	accuracy before deleting	distance before deleting	accuracy after deleting	distance after deleting
0	0.00%	1.3751	50.68%	1.1724
1	18.84%	1.2131	16.15%	1.1131
2	98.57%	0.7865	97.86%	0.9538
3	85.00%	1.0227	88.07%	0.9802
4	70.00%	1.1178	79.50%	1.0008
5	74.00%	1.1507	87.00%	0.9984
6	73.36%	1.1598	69.63%	1.1598
7	92.09%	0.9469	69.49%	1.1261
8	51.16%	1.1090	50.87%	1.1090
9	68.37%	1.1884	89.98%	1.0662
11	45.33%	1.0920	42.67%	1.0839
12	1.85%	1.5678	85.93%	0.9424
13	82.48%	1.1473	83.94%	1.1473
14	83.10%	1.0650	47.18%	1.2314
15	4.49%	1.3621	4.49%	1.3621
16	58.24%	1.1194	59.77%	1.1194
17	54.55%	1.2466	1.36%	1.3393
18	2.15%	1.5406	0.86%	1.5406
19	93.94%	1.1197	73.48%	1.1197
20	91.21%	1.0916	92.86%	1.0916
21	87.10%	1.1619	86.02%	1.1619
22	34.30%	1.1724	33.51%	1.1724
23	31.01%	1.1247	41.86%	1.1247
24	68.00%	0.9985	32.80%	1.0771
25	83.50%	1.1236	89.50%	1.0250
26	70.48%	1.1451	69.52%	1.1171
27	42.06%	1.2289	36.51%	1.2289
28	85.33%	0.9824	87.26%	0.9777
29	2.11%	1.3391	67.61%	1.1994
30	56.17%	1.1317	52.77%	1.1317
31	9.52%	1.2959	8.99%	1.2959
32	65.49%	1.1861	88.73%	1.0284
33	18.85%	1.1694	94.76%	0.8927
34	49.17%	1.1127	39.17%	1.1127
35	88.98%	1.1088	78.74%	1.1672
36	36.57%	1.2661	61.14%	1.1707
37	47.57%	1.1716	50.49%	1.1716
38	100%	0.8367	86.92%	1.1444
39	90.20%	1.1370	99.35%	0.8762
40	75.18%	1.0452	73.76%	1.0452
41	56.82%	1.2302	82.39%	1.1520
average	**57.83**%	**1.1556**	**62.08**%	**1.1185**

Qualitative Results. Next, I analyze the retrieval performance of my proposed model qualitatively in Figure 5.5. Some notable examples are as follows. Sketch query of cannon retrieves some examples of truck probably because both of them have wheels in common. Sketching guitar retrieves wine bottle because of the shape. Querying zebra, retrieves images like horse. In general, I observe that the wrongly retrieved candidates mostly have a closer visual with the queried ones. This effect is more prominent in DomainNet dataset, which may be due to the inter-class similarity of sketches between different classes

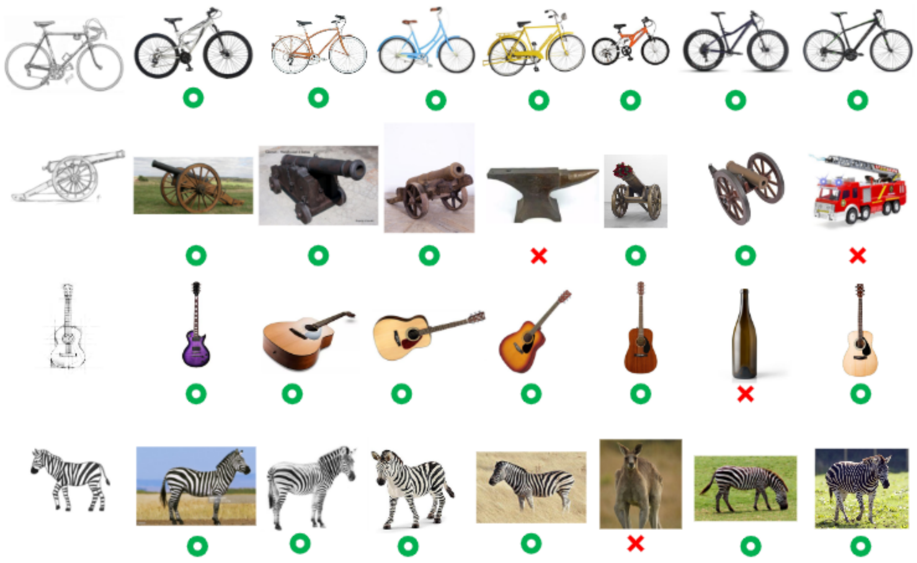


Figure 5.5: Top 7 image retrieval examples given a query sketch. All the examples correspond to a few-shot setting(20-shot). First two rows provides a retrieval result from Sketchy Dataset and last two rows shows the result of TU-Berlin Dataset.

5.5 Summary

I extend my single source domain adaptation algorithm to the multi-source scenarios. I estimate the means of the classes in the target domain in terms of the multi-source domain datasets. I pick a dataset with six domains and implement my multi-source domain adaptation method by choosing five domains as the source and one domain as target. It can be noted that increasing source domains, the improvement of retrieval performance converges around 4 source domains. This implies that the noise from different domains may be accumulated and seriously deteriorate the retrieval performance. The experiment also shows that my model can make use of different kinds of domains if they share the same class and I can simply improve my retrieval result by import new dataset.

Chapter 6

Conclusion and Future Work

This chapter concludes the thesis by synthesizing my contributions to enhancing generalization in Sketch-Based Image Retrieval through domain adaptation techniques. I reflect on the key findings, discuss limitations of the current approaches, and outline promising directions for future research.

6.1 Conclusion

This thesis has addressed the fundamental challenge of enhancing generalization in Sketch-Based Image Retrieval (SBIR) systems through the development of novel domain adaptation techniques. At its core, my work has established a comprehensive framework for quantitatively assessing generalization capabilities of deep networks, particularly in the context of SBIR, and has implemented this framework through innovative single-source and multi-source domain adaptation approaches.

The research journey began with the development of an empirical generalization metric that assesses both classification accuracy and data diversity handling capabilities—two critical factors for SBIR performance. Unlike traditional theoretical approaches to generalization, my framework provides a practical, quantifiable assessment method that offers reliable insights into model selection for SBIR applications. By identifying optimal trade-off points between model complexity, robustness, and generalization capacity, this metric system enables more informed selection of foundation models for SBIR.

Building upon this generalization framework, I introduced a single-source domain adaptation approach for SBIR that effectively bridges the domain gap between sketches and images. This method employs canonical correlation analysis to identify a shared subspace where source and target domains are maximally correlated, facilitating efficient knowledge transfer with minimal labeled examples in the target domain. The integration of low-rank matrix decomposition and sparse optimization techniques ensures computational efficiency while maintaining high adaptation fidelity—crucial considerations for practical SBIR applications.

I further extended my approach to multi-source domain adaptation scenarios, addressing the more complex yet practically valuable case where multiple diverse source domains are available. This extension employs online dictionary learning alongside canonical correlation analysis to effectively combine information from multiple sources while filtering out domain-specific noise. my experiments demonstrated that this approach significantly enhances SBIR performance across diverse domains, though with an observed threshold beyond which the integration of additional sources yields diminishing returns.

Throughout this work, I maintained a balance between theoretical advancement and practical applicability, ensuring that my methods not only contribute to the understanding of generalization but also provide implementable solutions for real-world SBIR systems. The domain adaptation algorithms introduced in this thesis demonstrate a crucial equilibrium between performance enhancement and computational efficiency, offering scalable solutions that can be readily deployed in practical applications.

In sum, this thesis has established a comprehensive framework for enhancing generalization in SBIR through domain adaptation, providing both the theoretical foundation for understanding generalization in this context and practical methodologies for implementing effective cross-domain retrieval systems. The integration of traditional machine learning techniques with modern deep learning approaches has yielded robust solutions that advance the state of the art in SBIR, particularly for scenarios with limited labeled data in target domains.

6.2 Limitations

Despite the advancements presented in this thesis, several limitations warrant acknowledgment and provide motivation for future research:

6.2.1 Methodological Limitations

Linear Approximation Constraints: my domain adaptation approaches primarily rely on linear approximations through canonical correlation analysis and low-rank matrix decomposition. While these methods offer computational efficiency, they may not capture the full complexity of non-linear relationships between sketches and images, potentially limiting performance in highly complex scenarios.

Sensitivity to Feature Representation: The effectiveness of my methods depends significantly on the quality of the initial feature representations. Poor feature extraction from either sketches or images can propagate through the domain adaptation process, limiting overall retrieval performance regardless of adaptation quality.

Optimization Challenges: The non-convex nature of my optimization objectives, particularly in the multi-source scenario, introduces sensitivity to initialization and may lead to convergence to local optima rather than global solutions.

6.2.2 Data and Evaluation Limitations

Dataset Biases: The benchmark datasets used for evaluation, while comprehensive, may not fully represent the diversity of real-world sketching styles and image types. This potential mismatch between evaluation data and practical application scenarios could affect the generalizability of my findings.

Categorical Retrieval Focus: my evaluation primarily focuses on categorical retrieval (retrieving images from the same category as the query sketch) rather than instance-level retrieval. This limitation restricts the assessment of fine-grained matching capabilities that might be required in specific applications.

Static Evaluation: my evaluation methodology utilizes pre-drawn sketches rather than real-time sketches created in interactive settings. This approach may not capture the temporal aspects and variability inherent in live sketching scenarios.

6.2.3 Application Limitations

Absence of Semantic Information: Unlike some contemporary approaches, my methods do not incorporate textual or semantic information beyond visual features. This absence may limit performance in scenarios where visual ambiguity could be resolved through semantic context.

Computational Requirements: Despite my efforts to ensure efficiency, the matrix decomposition and correlation analysis procedures still require significant computational resources for large-scale datasets, potentially limiting applicability in resource-constrained environments.

Domain Coverage: my approach has been validated on a limited set of visual domains. The generalizability to other modalities, such as 3D models, videos, or other multimedia formats, remains unexplored and represents a potential limitation in broader multimedia retrieval contexts.

6.3 Future Work

Building upon the foundations established in this thesis and addressing its limitations, several promising directions for future research emerge:

Deepening Generalization Framework Analysis: The generalization framework for SBIR could be further refined and explored through different network architectures or learning paradigms. Future research could investigate how unsupervised or self-supervised learning approaches affect generalization in SBIR contexts, potentially reducing reliance on labeled data while maintaining retrieval performance.

Investigating Extreme Generalization Scenarios: A valuable direction would be to rigorously test the limits of my DA-SBIR algorithms and generalization metrics under extreme conditions. This includes scenarios with extremely sparse data (e.g., single-shot learning), highly abstract or stylized sketches, or application to massive-scale datasets with millions of images and thousands of categories.

Incorporating Semantic Information: While my current methods focus exclusively on visual features, integrating semantic information could significantly enhance retrieval performance. Future work could explore how to effectively combine my domain adaptation approaches with text embeddings, attribute information, or knowledge graphs to create multimodal SBIR systems with improved generalization capabilities.

Non-linear Domain Adaptation Approaches: Extending my current linear approximation methods to non-linear domain adaptation techniques could capture more complex relationships between domains. This might involve integrating deep canonical correlation analysis or developing hybrid approaches that combine the computational efficiency of my current methods with the representational power of deep neural networks.

Dynamic and Interactive SBIR: Future research could explore the temporal dimension of sketching by incorporating real-time, stroke-by-stroke analysis into the retrieval process. This would address the static evaluation limitation and enable more interactive SBIR systems that provide feedback and refine results as sketches are being drawn.

Cross-Modal Retrieval Extension: The principles and techniques developed in this thesis could be extended to other cross-modal retrieval scenarios beyond sketch-to-image mapping. Potential applications include sketch-to-3D model retrieval, sketch-to-video retrieval, or integration with other input modalities like voice descriptions or text queries.

Adaptive Domain Weighting: For multi-source domain adaptation, developing methods that dynamically weight the contribution of each source domain based on its relevance to the target could further improve performance. This approach would address the diminishing returns observed when integrating multiple sources and could potentially enable more effective knowledge transfer from diverse domain collections.

User-Adaptive SBIR: A particularly promising direction is the development of SBIR systems that adapt to individual users' sketching styles over time. Such personalized systems would combine my domain adaptation techniques with online learning approaches to continuously refine retrieval performance based on user feedback and sketching patterns.

By pursuing these research directions, future work can address the current limitations while expanding the capabilities and applications of SBIR systems, ultimately bringing us closer to the goal of intuitive, robust, and broadly applicable sketch-based visual search technologies.

References

Akaho, S., 2006. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

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

Akata, Z., Reed, S., Walter, D., Lee, H. and Schiele, B., 2015b. Evaluation of output embeddings for fine-grained image classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2927–2936.

Al-Halah, Z., Tapaswi, M. and Stiefelhagen, R., 2016. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5975–5984.

Andrew, G., Arora, R., Bilmes, J. and Livescu, K., 2013. Deep canonical correlation analysis. *International conference on machine learning*, PMLR, 1247–1255.

Ao, S., Li, X. and Ling, C., 2017. Fast generalized distillation for semi-supervised domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Arora, S., Cohen, N. and Hazan, E., 2018. On the optimization of deep networks: Implicit acceleration by overparameterization. *International Conference on Machine Learning*, PMLR, 244–253.

Babenko, A., Slesarev, A., Chigorin, A. and Lempitsky, V., 2014. Neural codes for image retrieval. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 584–599.

Bach, F. R. and Jordan, M. I., 2002. Kernel independent component analysis. *Journal of machine learning research*, 3 (Jul), 1–48.

Barron, A. R. and Klusowski, J. M., 2019. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*.

Bartlett, P. L. and Mendelson, S., 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3 (Nov), 463–482.

Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8), 1798–1828.

Brutzkus, A., Globerson, A., Malach, E. and Shalev-Shwartz, S., 2017. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*.

Bubeck, S. and Sellke, M., 2023. A universal law of robustness via isoperimetry. *J. ACM*, 70 (2). URL `https://doi.org/10.1145/3578580`.

Chandar, S., Khapra, M. M., Larochelle, H. and Ravindran, B., 2016. Correlational neural networks. *Neural computation*, 28 (2), 257–285.

Chang, X., Xiang, T. and Hospedales, T. M., 2018. Scalable and effective deep cca via soft decorrelation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1488–1497.

Changpinyo, S., Chao, W.-L., Gong, B. and Sha, F., 2016. Synthesized classifiers for zero-shot learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5327–5336.

Chattopadhyay, R., Sun, Q., Fan, W., Davidson, I., Panchanathan, S. and Ye, J., 2012. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6 (4), 1–26.

Chen, Y., Zhu, X. and Gong, S., 2018. Semi-supervised deep learning with memory. *Proceedings of the European conference on computer vision (ECCV)*, 268–283.

Chopra, S., Balakrishnan, S. and Gopalan, R., 2013. Dlid: Deep learning for domain adaptation by interpolating between domains. *ICML workshop on challenges in representation learning*, Citeseer, volume 2.

Chu, B., Madhavan, V., Beijbom, O., Hoffman, J. and Darrell, T., 2016. Best practices for fine-tuning visual classifiers to new domains. *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, Springer, 435–442.

Chu, W.-S., De la Torre, F. and Cohn, J. F., 2013. Selective transfer machine for personalized facial action unit detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3515–3522.

Cicek, S., Fawzi, A. and Soatto, S., 2018. Saas: Speed as a supervisor for semi-supervised learning. *Proceedings of the European Conference on Computer Vision (ECCV)*, 149–163.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (1), 37–46.

Csurka, G., Chidlovskii, B. and Perronnin, F., 2015. Domain adaptation with a domain specific class means classifier. *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*, Springer, 32–46.

Csurka, G., Chidlowskii, B., Clinchant, S. and Michel, S., 2016. Unsupervised domain adaptation with regularized domain instance denoising. *European Conference on Computer Vision*, Springer, 458–466.

Daumé III, H., 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 248–255.

Dey, S., Riba, P., Dutta, A., Llados, J. and Song, Y.-Z., 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2179–2188.

Doersch, C., Gupta, A. and Efros, A. A., 2015. Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE international conference on computer vision*, 1422–1430.

Donahue, J., Hoffman, J., Rodner, E., Saenko, K. and Darrell, T., 2013. Semi-supervised domain adaptation with instance constraints. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 668–675.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T., 2014. Decaf: A deep convolutional activation feature for generic visual recognition. *International conference on machine learning*, PMLR, 647–655.

Dong, C., Li, W., Huo, J., Gu, Z. and Gao, Y., 2021. Learning task-aware local representations for few-shot learning. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 716–722.

Dorfer, M., Schlüter, J., Vall, A., Korzeniowski, F. and Widmer, G., 2018. End-to-end cross-modality retrieval with cca projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval*, 7 (2), 117–128.

Duan, L., Xu, D. and Tsang, I. W.-H., 2012. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on neural networks and learning systems*, 23 (3), 504–518.

Dutta, A. and Akata, Z., 2020. Semantically tied paired cycle consistency for any-shot sketch-based image retrieval. *International Journal of Computer Vision*, 128 (10), 2684–2703.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I. and Roy, D. M., 2020. In search of robust measures of generalization. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., NIPS '20.

Dziugaite, G. K. and Roy, D. M., 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Eitz, M., Hays, J. and Alexa, M., 2012. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31 (4), 1–10.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D. and Equitz, W., 1994. Efficient and effective querying by image content. *Journal of intelligent information systems*, 3 (3-4), 231–262.

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

Fu, Z., Xiang, T., Kodirov, E. and Gong, S., 2015. Zero-shot object recognition by semantic manifold distance. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2635–2644.

Gao, L., Song, J., Zou, F., Zhang, D. and Shao, J., 2015. Scalable multimedia retrieval by deep learning hashing with relative similarity learning. *Proceedings of the 23rd ACM international conference on Multimedia*, 903–906.

Garcia, V. and Bruna, J., 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

Gebelein, H., 1941. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21 (6), 364–379.

Glorot, X., Bordes, A. and Bengio, Y., 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the 28th international conference on machine learning (ICML-11)*, 513–520.

Golowich, N., Rakhlin, A. and Shamir, O., 2018. Size-independent sample complexity of neural networks. *Conference On Learning Theory*, PMLR, 297–299.

Gong, B., Shi, Y., Sha, F. and Grauman, K., 2012. Geodesic flow kernel for unsupervised domain adaptation. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2066–2073.

Gong, Y., Wang, L., Guo, R. and Lazebnik, S., 2014. Multi-scale orderless pooling of deep convolutional activation features. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, Springer, 392–407.

Gordo, A., Almazán, J., Revaud, J. and Larlus, D., 2016. Deep image retrieval: Learning global representations for image search. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, Springer, 241–257.

Hardoon, D. R., Szedmak, S. and Shawe-Taylor, J., 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16 (12), 2639–2664.

Hardt, M., Recht, B. and Singer, Y., 2016. Train faster, generalize better: Stability of stochastic gradient descent. *International conference on machine learning*, PMLR, 1225–1234.

Harvey, N., Liaw, C. and Mehrabian, A., 2017. Nearly-tight vc-dimension bounds for piecewise linear neural networks. *Conference on learning theory*, PMLR, 1064–1068.

Hirata, K. and Kato, T., 1992. Query by visual example: Content based image retrieval. *international conference on extending database technology*, Springer, 56–71.

Hirschfeld, H. O., 1935. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, volume 31, 520–524.

Hoffer, E. and Ailon, N., 2015. Deep metric learning using triplet network. *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, Springer, 84–92.

Hoffman, J., Rodner, E., Donahue, J., Darrell, T. and Saenko, K., 2013. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*.

Hotelling, H., 1992. Relations between two sets of variates. *Breakthroughs in statistics*, Springer, 162–190.

Hu, R. and Collomosse, J., 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117 (7), 790–806.

Jayaraman, D. and Grauman, K., 2014. Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems*, 27.

Jiang, Y., Foret, P., Yak, S., Roy, D. M., Mobahi, H., Dziugaite, G. K., Bengio, S., Gunasekar, S., Guyon, I. and Neyshabur, B., 2020. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*.

Jiang, Y., Krishnan, D., Mobahi, H. and Bengio, S., 2018. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. and Bengio, S., 2019. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

Kim, H., Park, J., Choi, Y. and Lee, J., 2023. Fantastic robustness measures: The secrets of robust generalization. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds., *Advances in Neural Information Processing Systems*, Curran Associates, Inc., volume 36, 48793–48818. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/98a5c0470e57d518ade4e56c6ee0b363-Paper-Conference.pdf.

Koch, G., Zemel, R., Salakhutdinov, R. et al., 2015. Siamese neural networks for one-shot image recognition. *ICML deep learning workshop*, Lille, volume 2, 0.

Kodirov, E., Xiang, T. and Gong, S., 2017. Semantic autoencoder for zero-shot learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3174–3183.

Krizhevsky, A., Hinton, G. et al., 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Laine, S. and Aila, T., 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Lampert, C. H., Nickisch, H. and Harmeling, S., 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36 (3), 453–465.

Leggetter, C. J. and Woodland, P. C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech & language*, 9 (2), 171–185.

Li, B., Jin, J., Zhong, H., Hopcroft, J. and Wang, L., 2022. Why robust generalization in deep learning is difficult: Perspective of expressive power. *Advances in Neural Information Processing Systems*, 35, 4370–4384.

Li, Y., Wang, D., Hu, H., Lin, Y. and Zhuang, Y., 2017. Zero-shot recognition using dual visual-semantic mapping paths. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3279–3287.

Lin, H.-W., Bao, H.-J. and Wang, G.-J., 2005. Totally positive bases and progressive iteration approximation. *Computers & Mathematics with Applications*, 50 (3-4), 575–586.

Lin, K., Lu, J., Chen, C.-S. and Zhou, J., 2016. Learning compact binary descriptors with unsupervised deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1183–1192.

Lin, K., Yang, H.-F., Hsiao, J.-H. and Chen, C.-S., 2015. Deep learning of binary hash codes for fast image retrieval. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 27–35.

Lindenbaum, O., Salhov, M., Averbuch, A. and Kluger, Y., 2021. L0-sparse canonical correlation analysis. *International Conference on Learning Representations*.

Liu, L., Shen, F., Shen, Y., Liu, X. and Shao, L., 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2862–2871.

Lopez-Paz, D., Sra, S., Smola, A., Ghahramani, Z. and Schölkopf, B., 2014. Randomized nonlinear component analysis. *International conference on machine learning*, PMLR, 1359–1367.

Lu, P., Huang, G., Lin, H., Yang, W., Guo, G. and Fu, Y., 2021. Domain-aware se network for sketch-based image retrieval with multiplicative euclidean margin softmax. *Proceedings of the 29th ACM International Conference on Multimedia*, 3418–3426.

Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11).

Mancini, M., Akata, Z., Ricci, E. and Caputo, B., 2020. Towards recognizing unseen categories in unseen domains. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, Springer, 466–483.

Mehrkanoon, S. and Suykens, J. A., 2017. Regularized semipaired kernel cca for domain adaptation. *IEEE transactions on neural networks and learning systems*, 29 (7), 3199–3213.

Melzer, T., Reiter, M. and Bischof, H., 2001. Nonlinear feature extraction using generalized canonical correlation analysis. *International Conference on Artificial Neural Networks*, Springer, 353–360.

Michaeli, T., Wang, W. and Livescu, K., 2016. Nonparametric canonical correlation analysis. *International conference on machine learning*, PMLR, 1967–1976.

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

Miyato, T., Maeda, S.-i., Koyama, M. and Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41 (8), 1979–1993.

Mou, W., Wang, L., Zhai, X. and Zheng, K., 2018. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *Conference on Learning Theory*, PMLR, 605–638.

Natekar, P. and Sharma, M., 2020. Representation based complexity measures for predicting generalization in deep learning.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. and Srebro, N., 2018. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*.

Niblack, C. W., Barber, R., Equitz, W., Flickner, M. D., Glasman, E. H., Petkovic, D., Yanker, P., Faloutsos, C. and Taubin, G., 1993. Qbic project: querying images by content, using color, texture, and shape. *Storage and retrieval for image and video databases*, International Society for Optics and Photonics, volume 1908, 173–187.

Oquab, M., Bottou, L., Laptev, I. and Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.

Pan, S. J., Tsang, I. W., Kwok, J. T. and Yang, Q., 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22 (2), 199–210.

Pati, Y. C., Rezaiifar, R. and Krishnaprasad, P. S., 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Proceedings of 27th Asilomar conference on signals, systems and computers*, IEEE, 40–44.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K. and Wang, B., 2019. Moment matching for multi-source domain adaptation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1406–1415.

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

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D., 2008. *Dataset shift in machine learning*. Mit Press.

Radenović, F., Tolias, G. and Chum, O., 2016. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 3–20.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International conference on machine learning*, PMLR, 8748–8763.

Rényi, A., 1959. On measures of dependence. *Acta mathematica hungarica*, 10 (3-4), 441–451.

Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., 2000. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10 (1-3), 19–41.

Romera-Paredes, B. and Torr, P., 2015. An embarrassingly simple approach to zero-shot learning. *International conference on machine learning*, PMLR, 2152–2161.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3), 211–252.

Saavedra, J. M., 2014. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). *2014 IEEE international conference on image processing (ICIP)*, IEEE, 2998–3002.

Saavedra, J. M., Barrios, J. M. and Orand, S., 2015. Sketch based image retrieval using learned keyshapes (lks). *BMVC*, volume 1, 7.

Saenko, K., Kulis, B., Fritz, M. and Darrell, T., 2010. Adapting visual category models to new domains. *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, Springer, 213–226.

Saito, K., Kim, D., Sclaroff, S., Darrell, T. and Saenko, K., 2019. Semi-supervised domain adaptation via minimax entropy. *Proceedings of the IEEE/CVF international conference on computer vision*, 8050–8058.

Sanae Lotfi, Y. K. T. R. M. G. A. W., Marc Finzi, 2023. Non-vacuous generalization bounds for large language models. *Proceedings of Workshop Mathematics of Modern Machine Learning (M3L) of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., Workshop of NIPS '23.

Sangkloy, P., Burnell, N., Ham, C. and Hays, J., 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35 (4), 1–12.

Saxena, S. and Verbeek, J., 2016. Heterogeneous face recognition with cnns. *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, Springer, 483–491.

Schonfeld, E., Ebrahimi, S., Sinha, S., Darrell, T. and Akata, Z., 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Schroff, F., Kalenichenko, D. and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Sharif Razavian, A., Azizpour, H., Sullivan, J. and Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 806–813.

Shawe-Taylor, J. and Williamson, R. C., 1997. A pac analysis of a bayesian estimator. *Proceedings of the tenth annual conference on Computational learning theory*, 2–9.

Shen, Y., Liu, L., Shen, F. and Shao, L., 2018. Zero-shot sketch-image hashing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3598–3607.

Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90 (2), 227–244.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socher, R., Ganjoo, M., Manning, C. D. and Ng, A., 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26.

Song, J., Yu, Q., Song, Y.-Z., Xiang, T. and Hospedales, T. M., 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. *Proceedings of the IEEE international conference on computer vision*, 5551–5560.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. and Kawanabe, M., 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20.

Sun, B., Feng, J. and Saenko, K., 2016. Return of frustratingly easy domain adaptation. *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tan, M. and Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, PMLR, 6105–6114.

Tarvainen, A. and Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Thanh-Tung, H. and Tran, T., 2020. Toward a generalization metric for deep generative models. *arXiv preprint arXiv:2011.00754*.

Tolias, G., Sicre, R. and Jégou, H., 2015. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*.

Valle-Pérez, G. and Louis, A. A., 2020. Generalization bounds for deep learning. *arXiv preprint arXiv:2012.04115*.

Vinod, H. D., 1976. Canonical ridge and econometrics of joint production. *Journal of econometrics*, 4 (2), 147–166.

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y. and Li, J., 2014. Deep learning for content-based image retrieval: A comprehensive study. *Proceedings of the 22nd ACM international conference on Multimedia*, 157–166.

Wang, F., Kang, L. and Li, Y., 2015a. Sketch-based 3d shape retrieval using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1875–1883.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B. and Wu, Y., 2014. Learning fine-grained image similarity with deep ranking. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1386–1393.

Wang, L., Wu, J., Huang, S.-L., Zheng, L., Xu, X., Zhang, L. and Huang, J., 2019a. An efficient approach to informative feature extraction from multimodal data. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5281–5288.

Wang, Q., Li, W. and Gool, L. V., 2019b. Semi-supervised learning by augmented distribution alignment. *Proceedings of the IEEE/CVF international conference on computer vision*, 1466–1475.

Wang, W., Arora, R., Livescu, K. and Bilmes, J., 2015b. On deep multi-view representation learning. *International conference on machine learning*, PMLR, 1083–1092.

Wang, X. and Gupta, A., 2015. Unsupervised learning of visual representations using videos. *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Wang, Z., Wang, H., Yan, J., Wu, A. and Deng, C., 2021. Domain-smoothing network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2106.11841*.

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

Xian, Y., Lampert, C. H., Schiele, B. and Akata, Z., 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41 (9), 2251–2265.

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

Xu, R., Chen, Z., Zuo, W., Yan, J. and Lin, L., 2018. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3964–3973.

Yan, F. and Mikolajczyk, K., 2015. Deep correlation for matching images and text. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3441–3450.

Yang, J., Yan, R. and Hauptmann, A. G., 2007. Cross-domain video concept detection using adaptive svms. *Proceedings of the 15th ACM international conference on Multimedia*, 188–197.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. and Chaudhuri, K., 2020. A closer look at accuracy vs. robustness. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., NIPS '20.

Yao, T., Pan, Y., Ngo, C.-W., Li, H. and Mei, T., 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2142–2150.

Yelamarthi, S. K., Reddy, S. K., Mishra, A. and Mittal, A., 2018. A zero-shot framework for sketch based image retrieval. *Proceedings of the European Conference on Computer Vision (ECCV)*, 300–317.

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H., 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Yu, Q., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T. M. and Loy, C.-C., 2016. Sketch me that shoe. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 799–807.

Zeiler, M. D. and Fergus, R., 2014. Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 818–833.

Zhang, B., Cai, T., Lu, Z., He, D. and Wang, L., 2021a. Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons. *International Conference on Machine Learning*. URL `https://api.semanticscholar.org/CorpusID:235185401`.

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O., 2021b. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64 (3), 107–115.

Zhang, H., Liu, S., Zhang, C., Ren, W., Wang, R. and Cao, X., 2016. Sketchnet: Sketch classification with web images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1105–1113.

Zhang, J., Shen, F., Liu, L., Zhu, F., Yu, M., Shao, L., Shen, H. T. and Van Gool, L., 2018. Generative domain-migration hashing for sketch-to-image retrieval. *Proceedings of the European conference on computer vision (ECCV)*, 297–314.

Zhang, Z. and Saligrama, V., 2015. Zero-shot learning via semantic similarity embedding. *Proceedings of the IEEE international conference on computer vision*, 4166–4174.

Zhang, Z. and Saligrama, V., 2016. Zero-shot learning via joint latent similarity embedding. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6034–6042.