國立臺灣大學電機資訊學院電信工程學研究所

博士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Doctoral Dissertation

知識遷移於視覺理解

Visual Understanding with Knowledge Transfer

楊福恩

Fu-En Yang

指導教授：王鈺強 博士

Advisor: Yu-Chiang Frank Wang, Ph.D.

中華民國 112 年 7 月

July, 2023

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

國立臺灣大學博士學位論文
口試委員會審定書
PhD DISSERTATION ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

（論文中文題目）(Chinese title of PhD dissertation)

知識遷移於視覺理解

（論文英文題目）(English title of PhD dissertation)

Visual Understanding with Knowledge Transfer

本論文係 楊福恩 (姓名) F07942077（學號）在國立臺灣大學 電信工程學研究所 (系/所/學位學程)完成之博士學位論文，於民國 112 年 7 月 12 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department / Institute of Communication Engineering on 12 (date) 7 (month) 2023 (year) have examined a PhD dissertation entitled above presented by Fu-En Yang (name) F07942077 (student ID) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

王鈺強 （指導教授 Advisor）	陳煥宗	邱維辰
陳祝嵩	李宏毅	莊永裕
孫紹華	孫民	

系主任/所長 Director: 周錫增

doi:10.6342/NTU202301924

致謝

本論文以及博士班學業的完成，首先感謝我的指導老師王鈺強教授，在研究上的指導與提點，從研究主題的發想，找到重要且值得解決的研究問題，到方法的設計，如何有效地解決問題並且設計實驗做驗證，到最後論文寫作時的邏輯脈絡和組織架構，能夠清楚地呈現出研究中的想法概念，這些訓練都讓我有很大的成長和經驗累積，相信會是未來受用無窮的實力。也非常感謝老師這幾年的支持和鼓勵，即便在投稿不順利時，也能及時調整，繼續向前邁進。很榮幸能成為 VLL 的一份子。

在這裡也要特別感謝研究過程中一起努力的好夥伴。感謝景程在專題生時就願意一起合作，雖然早期做研究時經驗還不足，但一起討論想法，跟寫作投稿的過程，真的是非常難忘，感謝你強大的實作能力，讓我們早期兩篇論文得以順利完成並發表。也感謝元顥在投稿時的強力支援，很懷念一起在實驗室聊天討論，還有當助教的時光。感謝元嘉和如芸，在當時因為疫情三級警戒時，仍然願意透過線上的方式，和我討論 NeurIPS 的實驗與投稿，給我很大的鼓勵，協助完成整個草稿和打 rebuttal，讓我在這次的投稿過程學到很多，感謝你們願意幫忙，才有這一篇論文的完成和順利發表。也感謝實驗室助理芳茹，感謝你總是細心地幫忙打點好各樣大小事，也總是能幫我撥開思緒中的五里霧，感謝有你，是實驗室的精神支柱。也謝謝 VLL 的大家，很開心能跟大家一起討論各種有趣的想法，腦力激盪可能的解法和方向，很感激一起走過研究所期間的點點滴滴，萬分感謝。

最後要感謝我的家人，無條件地支持我，讓我能夠沒有顧慮，在研究上能專心持續往前，同時也是我最大的避風港，包容我的一切。感謝文馨帶給我的勇氣與支持，讓我更有信心，陪我一起走過無論是順利或是顛頗的時刻，謝謝你總是傾聽也給我支持和鼓勵，許多地體貼包容，一起面對未來每一件大小事。感謝神一路的帶領，因為有信仰和教會，讓我在面對許多事情時，能用不同的眼光去面對和看待，不僅僅在自己的領域中精進，也期待自己做光做鹽，成為別人的幫助。

2023.07.24 楊福恩

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

中文摘要

深度學習的進步得益於大規模且精細蒐集的數據資料集。然而，這些數據集通常基於一個假設，即訓練和測試資料是共享相同的分佈。但在實際的應用場景，特別是在計算機視覺領域中，這樣的假設往往很難成立，在圖像領域分佈或是語義類別通常有所差異。由於這些資料分佈的不同，對特定分佈進行訓練的深度神經網絡在不同的資料分佈數據上往往表現不佳。在本論文中，我們的目標是透過遷移學習，以實現在不同的圖像領域分佈或語義類別之間進行知識的遷移。

在本論文中，我們首先解決圖像風格的知識轉移問題。我們提出了一個特徵解耦框架，實現跨多個圖像領域和多樣化的風格轉移。接著，我們研究語義類別的知識轉移，透過利用類別內觀察到的差異來完成零樣本圖像識別這一具有挑戰性的任務。為了讓訓練模型能更好地處理落在源域分佈之外的數據，我們提出了一種用於領域泛化的對抗性教師–學生表示學習框架。最後，我們轉向分佈式學習的場景，用以達成在特定應用場景，例如醫療上的隱私保護要求。為了解決這個問題，我們設計了一種針對特定數據領域的提示生成框架，來允許高效並且個性化的聯邦學習。通過實驗的分析與結果，本論文中提出的方法的有效性得以驗證。

關鍵字：深度學習、電腦視覺、遷移學習、風格轉換、零樣本學習、領域泛化、聯邦學習

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

Abstract

Recent progress in deep learning owes a lot to large-scale, curated datasets. However, these datasets typically operate on the assumption that training and test data share the same distribution. This is not always the case in real-world scenarios, particularly in the field of computer vision, where discrepancies in data domains or semantic categories are common. Due to these distribution gaps, deep neural networks trained on a specific distribution can struggle to perform in a different domain. In this thesis, we aim at advancing transfer learning to enable the transfer of knowledge across distinct data domains or semantic classes. Specifically, we first address knowledge transfer for image styles. We propose a feature disentanglement framework that facilitates multi-domain and multi-modal style transfer. Next, we examine knowledge transfer for semantic categories, focusing on the challenging task of zero-shot image recognition by leveraging intra-class variations. With the goal of enabling the trained model to handle data that falls outside the source distribution, we propose an Adversarial Teacher-Student Representation Learning framework for domain generalization. Finally, we transition to a decentralized learning paradigm, accommodating the privacy-preserving requirements of certain applications, such as healthcare. To tackle this, we devise a client-specific prompt generation framework to allow efficient, personalized federated learning. Through the comprehensive analysis and results, the effectiveness of the methods presented in this thesis could be successfully confirmed.

Keywords: deep learning, computer vision, transfer learning, style transfer, zero-shot learning, domain generalization, federated learning.

doi:10.6342/NTU202301924

Contents

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

List of Figures

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

List of Tables

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

Chapter 1

Knowledge Transfer for Image Styles

Learning interpretable data representation has been an active research topic in deep learning and computer vision. While representation disentanglement is an effective technique for addressing this task, existing works cannot easily handle the problems in which manipulating and recognizing data across multiple domains are desirable. In this thesis, we present a unified network architecture of Multi-domain and Multi-modal Representation Disentangler, with the goal of learning domain-invariant content representation with the associated domain-specific representation observed. By advancing adversarial learning and disentanglement techniques, the proposed model is able to perform continuous image manipulation across data domains with multiple modalities. More importantly, the resulting domain-invariant feature representation can be applied for unsupervised domain adaptation. Finally, our quantitative and qualitative results would confirm the effectiveness and robustness of the proposed model over state-of-the-art methods on the above tasks.

1.1 Introduction

Recent advances in deep learning have shown promising progresses in the areas of computer vision and machine learning. In particular, visual analysis and synthesis across data domains attract the attention from researchers in these fields. For

doi:10.6342/NTU202301924

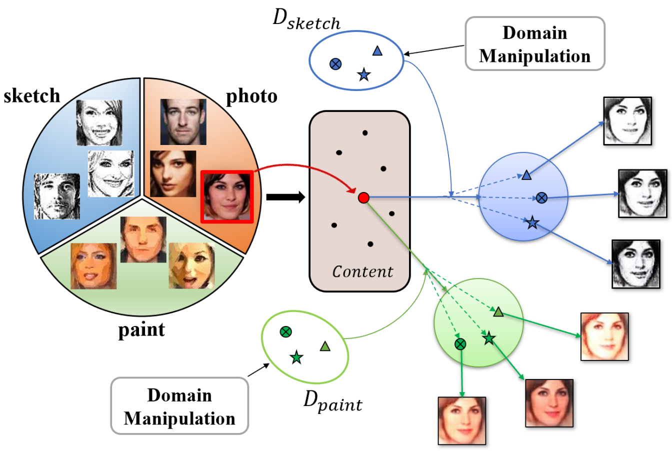


Figure 1.1: Illustration of multi-domain and multi-modal representation disentanglement. Given an input (in red bounding box) and images in multiple domains (e.g., styles), we derive representations for describing domain-invariant and domain-specific information, while images can be manipulated and recovered in different domains with sufficient diversity. Note that D_{sketch} and D_{paint} denote domain-specific spaces for *sketch* and *paint* images, respectively.

example, style transfer [4, 5, 6, 7, 8], image-to-image translation [9, 10, 11, 12, 13], and cross-domain visual classification (or domain adaptation) [14, 15, 16, 17, 18] can all be viewed as the associated applications.

To address the above tasks, previous works typically either learn a deterministic (i.e., unimodal) mapping from one data domain to another, or to embed desirable information into the resulting latent space to derive the data representation. The technique of representation disentanglement [19, 20, 21] particularly observes and manipulates specific feature attributes of interest, which has also been applied in the above tasks. Thus, one can view the attributes of interest as the meaningful factors inherent in image data, and further synthesize preferable outputs accordingly. For instance, one can manipulate the *style* attributes of the disentangled latent feature to achieve style transfer and image-to-image translation (e.g., photo $\leftrightarrow$ sketch [22]).

In practice, adaptation or translation between data domains needs to exhibit multi-modal diversity. That is, a single input instance may correspond to di-

doi:10.6342/NTU202301924

verse possible outputs, associated with the same attribute of interest (e.g., image style). Even with the promising models based on generative adversarial networks (GANs) [23], one might encounter mode collapse problems and fail to produce multi-modal outputs. Recently, MUNIT [24] and DRIT [25] utilize the disentangled representation for multi-modal translation, achieved by decomposing the latent feature into disjoint features to describe content and style information. While these models manipulate the style feature to synthesize diverse outputs, they cannot be easily extended to handle the image manipulation among multiple (i.e., more than two) domains due to their network architecture designs.

Table 1.1: Comparisons with recent works on image translation and manipulation.

	Unpaired data	Bidirectional translation	Shared representation	Feature disentanglement	Unified structure	Multiple domains	Multi-modal translation	Unsupervised domain adaptation
Pix2Pix [9]	-	-	-	-	-	-	-	-
CycleGAN [11]	√	√	-	-	-	-	-	-
StarGAN [26]	√	√	-	-	-	-	-	-
DTN [10]	√	-	√	-	-	-	-	√
CyCADA [18]	√	√	-	-	-	-	-	√
UNIT [13]	√	√	√	-	-	-	-	√
E-CDRD [22]	√	√	√	√	-	√	-	√
BicycleGAN [27]	-	√	√	√	-	-	√	-
CDDN [28]	-	√	√	√	-	-	√	-
MUNIT [24]	√	√	√	√	-	-	√	-
DRIT [25]	√	√	√	√	-	-	√	√
UFDN [1]	√	√	√	√	√	√	-	√
M^{2}RD (Ours)	√	√	√	√	√	√	√	√

In this thesis, we propose a unified framework of *Multi-domain and Multi-modal Representation Disentangler* (M^2RD) for cross-domain image synthesis and classification, with the ability to manipulate image data with the particular attribute of interest while exhibiting sufficient diversity, as illustrated in Fig. 1.1. Without collecting pairwise image data across domains, our model encodes image data into a domain-invariant and specific latent feature spaces. While the former observes content information from the input data, the latter exhibits multi-modal diversity during cross-domain image translation. In the experiments, we not only show that our model is able to perform image manipulation, but we further verify that derived domain-invariant content features can be applied to the task of unsuper-

doi:10.6342/NTU202301924

vised domain adaptation. With both qualitative and quantitative results provided, the effectiveness and robustness of our model can be successfully confirmed.

We now highlight the contributions as follows:

- Our proposed deep learning model is able to factorize latent image representations into disjoint features describing domain-invariant and specific information.

- Our network uniquely integrates adversarial learning, representation disentanglement, and generative modules in a unified architecture.

- Our derived domain-invariant feature representation allows unsupervised domain adaptation, while the domain-specific feature enables multi-modal image manipulation across multiple data domains.

1.2 Related Works

Representation Disentanglement. Aims at learning interpretable data representations ([19, 20, 21, 29, 30, 31, 32, 33]), Chen *et al.* [19] proposed InfoGAN to maximize the mutual information between the latent features and generated images, which realizes representation disentanglement in an unsupervised way. Similarly, Higgins *et al.* [20] introduced β-VAE which derives such representations by adding an adjustable hyperparameter to a variational auto-encoder (VAE) [34], balancing the latent channel capacity and the independence constraints. Tulyakov *et al.* [29] presented MoCoGAN to learn motion and content decomposition for video generation. Although the above methods realize representation disentanglement without label supervision, one cannot manipulate the latent factors directly since the semantic meanings behind the disentangled factors cannot be explicitly obtained. Thus, Odena *et al.* [21] augmented GANs with an auxiliary classifier, allowing image outputs to be conditioned on the desirable latent factors. Furthermore, Peng *et al.* [32] applied reconstruction-based disentanglement and self-supervision to

doi:10.6342/NTU202301924

guarantee completely decoupling of latent factors, which benefits pose-invariant face recognition. Tran *et al.* [30], and Liu *et al.* [31] proposed DR-GAN, and MTAN, which derived pose-invariant feature via disentanglement technique and adversarial learning to facilitate the face recognition. Tian *et al.* [33] employed GAN and cycle-consistency for disentangling latent features in multi-view image manipulation. Despite significant progresses, most existing works only focus on producing such representations from a single data domain.

Image-to-Image Translation. To convert images from one style to another, Isola *et al.* [9] chose to observe pairs of images for learning GAN-based models. Taigman *et al.* [10] presented Domain Transfer Network (DTN) to performed such tasks by employing feature consistency across domains. Without observing cross-domain image pairs, Zhu *et al.* [11] learned the bidirectional domain mappings in pixel space with a cycle consistency loss; similar ideas were also applied by [35] and [36]. Coupled GAN (CoGAN) [12] binds high-level information between two data domains for learning the joint distribution. UNIT [13] is extended from CoGAN, which integrates VAE and GAN to achieve image translation by mapping the data between two domains to the same latent space. While the above methods produce promising results, they cannot provide diverse outputs due to their model designs or issues like model collapse.

For multi-modal translation, Zhu *et al.* [27] observed pairs of images for deriving bijection mapping between the latent and output spaces. Gonzalez-Garcia *et al.* [28] decomposed the paired inputs into disjoint shared and exclusive parts to perform diverse image-to-image translation between two domains. Recently, Huang *et al.* [24] and Lee *et al.* [25] concurrently proposed MUNIT and DRIT respectively. MUNIT and DRIT both factorize the latent representation into a domain-invariant content feature and a domain-specific style feature from unpaired data. However, their model designs limit the use of data across multiple domains.

Cross-Domain Image Manipulation. In addition to image-to-image translation, several recent works [22, 37, 38, 1] further address image synthesis tasks

doi:10.6342/NTU202301924

with the ability of manipulating the attributes of interest. For example, Liu *et al.* [22] considered cross-domain disentangled representation with supervision from single-domain data which aims to manipulate the desirable attributes across different domains. However, they can only deal with a pair of data domains using the proposed model. To handle such tasks across multiple domains, Choi *et al.* [37], He *et al.* [38], and Liu *et al.* [1] proposed StarGAN, AttGAN, and UFDN respectively, which all perform multi-domain image-to-image translation by manipulating the domain label directly. Although StarGAN allows training of multiple domains simultaneously by the unified model structure, it does not exhibit ability in disentangling desirable latent representation. Nevertheless, while the above models are able to manipulate face images, our model further allows one to perform image-to-image translation on a variety of images including images of faces and natrual sccenes. Most importantly, all of them cannot allow multi-modal outputs, which might not be desirable for practical uses.

Unsupervised Domain Adaptation (UDA). Domain adaptation [39, 40, 14, 15, 16, 18] addresses the same learning tasks across domains, with the goal of eliminating the domain shift (i.e., dataset bias). And, unsupervised domain adaptation (UDA) specifically deals with the scenario in which no label supervision is available during training in the target domain. For instance, GAKT [39] applied an adaptive graph to transfer discriminative information from a labeled source to an unlabeled target domain. Also, Ding *et al.* [40] integrated low-rank coding with deep neural network for preserving global structures across source and target, to achieve more effective knowledge transfer. Recently, several GAN-based methods have been proposed for UDA. For example, Ganin *et al.* [14] introduced a Domain Adversarial Neural Network (DANN) framework which contains a domain classifier with its gradient reversal layer serving as a domain-invariant feature extractor. Tzeng *et al.* [15] adapted feature extractors and classifier of source and target domains by domain adversarial learning strategies to tackle UDA. Bousmalis *et al.* [16] utilized the decomposed representations to produce domain-invariant fea-

doi:10.6342/NTU202301924

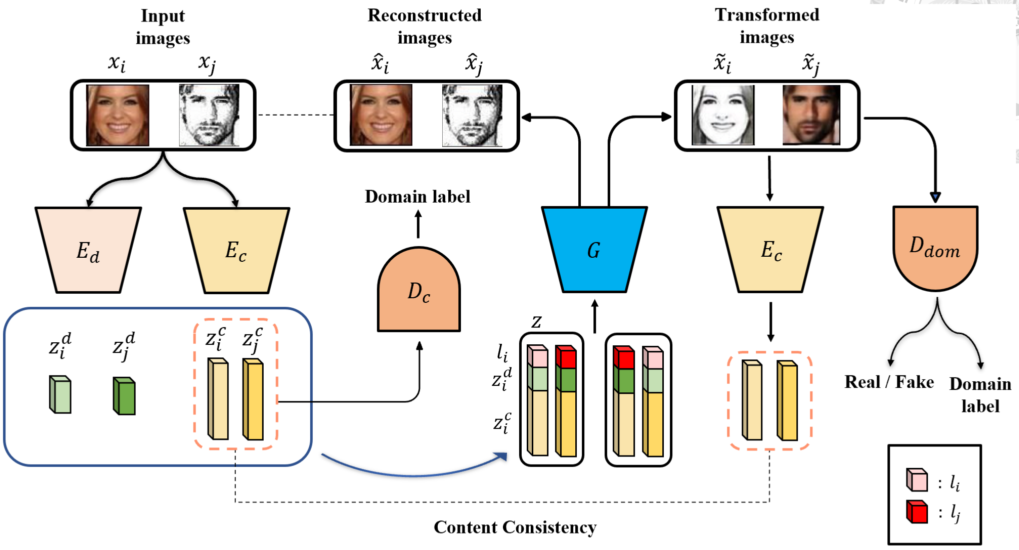


Figure 1.2: The network architecture of our Multi-domain and Multi-modal Representation Disentangler (M^{2}RD), which consists of two modules: 1) Representation disentangler, composed of a content encoder E_c, a domain encoder E_d, and a content discriminator D_c, and 2) Multi-domain and Multi-modal GAN consisting of a generator G, and a domain discriminator D_{dom} with an auxiliary domain classifier. Note that z^c, z^d denote the domain-invariant and specific features extracted from different domains respectively. Together with a domain code l, the final feature representation $z = [z^c, z^d, l]$ can be utilized for cross-domain and multi-modal image manipulation.

tures to facilitate cross-domain classification. Hoffman *et al.* [18] further extended CycleGAN [11] and applied adversarial learning and cycle-consistency for both feature and pixel-level adaptation.

Nevertheless, the above models typically do not exhibit abilities in disentangling particular image attributes, nor to manipulate image outputs across domains with multi-modal diversity. In Table 1.1, we compare our proposed model with recent deep learning methods in the aforementioned topics.

doi:10.6342/NTU202301924

1.3 Multi-domain and Multi-modal Representation Disentangler

1.3.1 Notation and Model Overview

Given an image set $\{\mathcal{X}_i\}_{i=1}^N$ across N distinct domains, our M^{2}RD jointly learns a domain-invariant content feature $\{z_i^c\}_{i=1}^N$ and domain-specific feature $\{z_i^d\}_{i=1}^N$ from the input image $x_i \in \mathcal{X}_i$, and then utilize discrete domain code $\{l_i\}_{i=1}^N$ to further exploit the domain information in the latent space. We note that the domain code l_i can be implemented by a one-hot vector, a real-value vector, or even concatenation of multiple one-hot vectors, which describes the domain of interest.

As illustrated in Fig. 1.2, our framework consists of two network modules. First, we have a representation disentangler with a content discriminator. This module contains a *content encoder* E_c and a *domain encoder* E_d, which are shared by input data across different domains. By advancing adversarial learning strategies, this disentangler module allows us to derive domain-invariant and specific features. The former provides the content information of the input data disregarding of its domain of origin, while the latter describes the domain of interest, which allows multi-modal manipulation as described later.

On the other hand, we have a Multi-domain and Multi-modal Generative Adversarial Networks as the second network module in Fig. 1.2, which includes a generator G and a domain discriminator D_{dom}, while the same content encoder E_c is deployed to observe *content consistency*. With the observed domain-invariant content feature z^c, this module performs both multi-domain and multi-modal image translation by manipulating the derived domain-specific feature z^d and the domain code l. The details of our proposed network will be discussed in the following subsections.

doi:10.6342/NTU202301924

1.3.2 Representation Disentangler

As illustrated in Fig. 1.2, our proposed network encodes cross-domain image inputs using shared content encoder E_c and domain encoder E_d. To enable the encoded content features to be domain-invariant, we apply a content discriminator D_c to eliminate the domain differences between the resulting features inspired by [14]. In other words, we have D_c aim to correctly produce domain code prediction $\hat{l}$ from the encoded content features z_i^c. Thus, the objective function of this content discriminator $\mathcal{L}_{adv}^{D_c}$ is derived as follows:

$$\mathcal{L}_{adv}^{D_c} = \mathbb{E}[\log(P(\hat{l} = l_i | E_c(x_i)))], \tag{1.1}$$

where P is the probability distribution over domains $\hat{l}$, which is calculated by the content discriminator D_c.

With the above design, our content encoder E_c would be able to extract the domain-invariant content features from input data, which are observed across multiple domains. As a result, the objective function of the encoder E_c is to maximize the cross-entropy of the content discriminator:

$$\mathcal{L}_{adv}^{E_c} = -\mathcal{L}_{adv}^{D_c} = -\mathbb{E}[\log(P(\hat{l} = l_i | E_c(x_i)))]. \tag{1.2}$$

Finally, in order to learn a joint and continuous representation for cross-domain data, and further perform stochastic sampling in the testing phase, we enforce the *Kullback-Leibler* divergence for our generative network model. This encourages the domain-specific feature z^d to fit a prior Gaussian distribution $N(0, I)$. Thus, the objective $\mathcal{L}_{KL}$ is calculated as:

$$\mathcal{L}_{KL} = \mathbb{E}[KL(E_d(x_i) || N(0, I))] \tag{1.3}$$

We note that, derivation of the above domain-invariant content representation is the reason why we can apply such features for unsupervised domain adaptation (UDA), which desires a common feature representation shared by different domains for adaptation purposes. With the above network design, we enforce the derived

doi:10.6342/NTU202301924

content features z^c does not contain any domain information, and thus the domain shift can be properly suppressed. As a result, we can simply deploy an extra classifier based on z^c if the UDA is of interest. To be more precise, the objective $\mathcal{L}_{cla}$ for this added UDA classifier can be expressed as follow:

$$\mathcal{L}_{cla} = -\sum_{k=1}^{N_{src}} y_k^{src} \cdot \log \tilde{y}_k^{src}. \tag{1.4}$$

where $\tilde{y}_k^{src}$ is the predicted output from the k-th labeled source input, and y_k^{src} is the ground truth label.

1.3.3 Multi-domain and Multi-modal GAN

Once the domain-invariant feature z^c and the domain-specific ones z^d are observed, the second module in our proposed architecture performs multi-domain and multi-modal image translation (i.e., cross-domain image manipulation with multi-modal diversity). We now discuss how these two tasks are jointly performed.

Similar to most existing image translation works, we perform image synthesis by combining the derived content feature z^c with the domain feature z^d. Extended from AC-GAN [21], we additionally assign the domain code l into the above feature representation to form the final feature representation $z = [z^c, z^d, l]$, followed by the decoding process.

Recall that, the representation z^d is learned to describe domain-only information, while such representation is shared by cross-domain data inputs. Thus, with the sampling strategies noted in Section 1.3.2, we will be able to reconstruct the image output and exhibit multi-modal diversity. In other words, *within-domain variants* of the recovered output associated with the same content feature z^c can be produced via sampling z^d. And, the above domain code l is added to ensure that the output image is recovered at or translated into the domain of interest. This is how our proposed model differs from existing image translation or disentanglement works.

With the above explanations, we now define the object functions applied in

doi:10.6342/NTU202301924

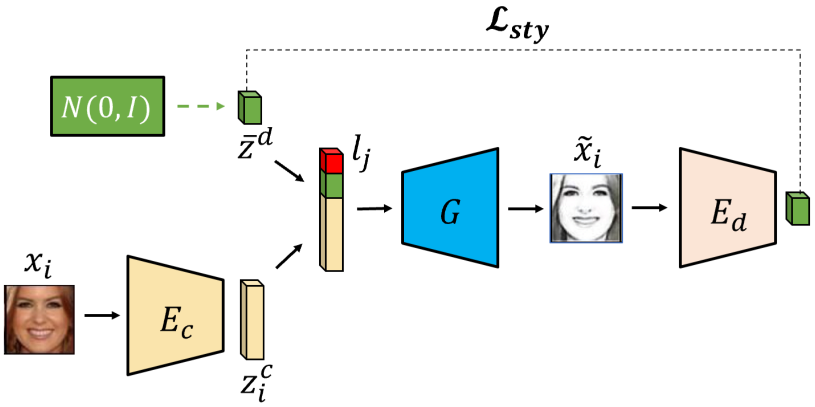


Figure 1.3: In addition to the architecture described in Fig. 1.2, we further apply the objective function $\mathcal{L}_{sty}$ to enforce the reconstruction on the domain-specific feature. More details can be found in Section 1.3.3

this network module. First, for image recovery guarantees, we calculate the reconstruction loss $\mathcal{L}_{rec}$ for the reconstructed image $\hat{x}_i$:

$$\mathcal{L}_{rec} = ||x_i - \hat{x}_i||_1, \tag{1.5}$$

Note that x_i is the (ground truth) input, and $\hat{x}_i = G([z_i^c, z_i^d, l_i])$.

Inspired by DTN [10], we further preserve the content consistency between translated images $\tilde{x}_i$ and input image x_i. Thus, an objective function $\mathcal{L}_{con}$ based on the same content encoder E_c is introduced in the feature level, which can be formulated as:

$$\mathcal{L}_{con} = ||E_c(x_i) - E_c(\tilde{x}_i)||_2, \tag{1.6}$$

where $\tilde{x}_i = G([z_i^c, z_j^d, l_j]),\ \ i \neq j$.

Also, similar to DRIT [25], we utilize style regression loss to enforce the reconstruction on the domain-specific feature, as illustrated in Fig. 1.3, with the objective $\mathcal{L}_{sty}$ expressed as:

$$\mathcal{L}_{sty} = ||E_d(G([z_i^c, \bar{z}^d, l_j]))) - \bar{z}^d||_2, \tag{1.7}$$

where $\bar{z}^d$ is sampled from a prior Gaussian distribution $N(0, I)$.

doi:10.6342/NTU202301924

However, when manipulating images across domains using the above network module, there is no guarantee that the output image $\tilde{x}_i$ would properly satisfy the domain information based on the domain code l inserted. Thus, as a part of the AC-GAN extension, we deploy a domain discriminator D_{dom} in Fig. 1.2 which performs multi-task learning for combining adversarial learning with an auxiliary domain classification task.

To be more precise, this discriminator not only determines the authenticity of the output images, it also classifies its domain label output to enforce the ability of the introduced domain code for domain disentanglement. Thus, the objective functions of this domain discriminator D_{dom} and generator G are calculated as follows:

$$\mathcal{L}_{adv}^{D_{dom}} = \mathbb{E}[\log(D_{dom}(\tilde{x}_i))] + \mathbb{E}[\log(1 - D_{dom}(x_i))], \tag{1.8}$$

$$\mathcal{L}_{aux}^{D_{dom}} = \mathbb{E}[\log(P(\bar{l} = l_j|\tilde{x}_i))] + \mathbb{E}[\log(P(\bar{l} = l_i|x_i))], \tag{1.9}$$

$$\mathcal{L}_{adv}^{G} = -\mathbb{E}[\log(D_{dom}(\tilde{x}_i))], \tag{1.10}$$

where $\bar{l}$ denotes the prediction output of D_{dom}. We note that the objective $\mathcal{L}_{aux}^{D_{dom}}$ aims at maximizing the mutual information between the domain code and the translated image [19].

1.3.4 Full Objectives

In summary, the full objective function $\mathcal{L}$ of our model can be summarized below:

$$\begin{aligned}\mathcal{L} = {} & \lambda_1 \mathcal{L}_{adv}^{D_c} + \lambda_1 \mathcal{L}_{adv}^{E_c} \\ & + \lambda_2(\mathcal{L}_{adv}^{D_{dom}} + \mathcal{L}_{aux}^{D_{dom}}) + \lambda_2 \mathcal{L}_{adv}^{G} \\ & + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{con}\mathcal{L}_{con} \\ & + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{sty}\mathcal{L}_{sty} + \lambda_{cla}\mathcal{L}_{cla},\end{aligned} \tag{1.11}$$

where the hyperparameters λ regularize each loss term. Nevertheless, we fix the values of λ for each dataset, and do not fine-tune them for each input instance.

doi:10.6342/NTU202301924

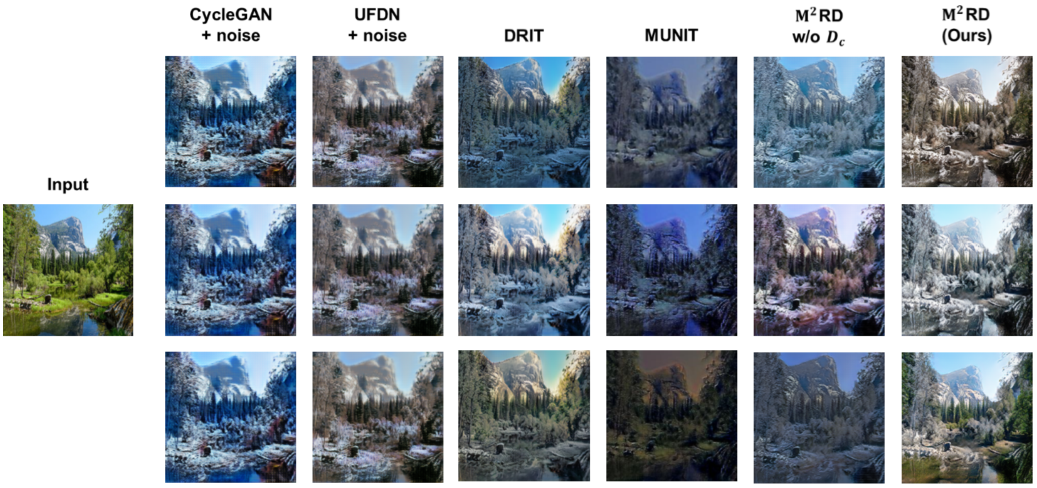


Figure 1.4: Example results of our multi-modal image translations and the comparison with the existing image-to-image translation methods. We observe that our model is able to generate high-quality images with meaningful diversity.

To train our model, we alternatively update content encoder E_c, domain encoder E_d, generator G, content discriminator D_c, and domain discriminator D_{dom} via the following gradients:

$$\begin{aligned}
\theta_{E_c} &\xleftarrow{+} -\Delta_{\theta_{E_c}}(\mathcal{L}_{rec} + \mathcal{L}_{adv}^{E_c} + \mathcal{L}_{con} + \mathcal{L}_{sty}) \\
\theta_{E_d} &\xleftarrow{+} -\Delta_{\theta_{E_d}}(\mathcal{L}_{rec} + \mathcal{L}_{KL}) \\
\theta_{G} &\xleftarrow{+} -\Delta_{\theta_{G}}(\mathcal{L}_{rec} + \mathcal{L}_{con} + \mathcal{L}_{KL} + \mathcal{L}_{adv}^{G} + \mathcal{L}_{aux}^{D_{dom}} + \mathcal{L}_{sty}) \qquad (1.12) \\
\theta_{D_c} &\xleftarrow{+} -\Delta_{\theta_{D_c}}(\mathcal{L}_{adv}^{D_c}) \\
\theta_{D_{dom}} &\xleftarrow{+} -\Delta_{\theta_{D_{com}}}(\mathcal{L}_{adv}^{D_{dom}} + \mathcal{L}_{aux}^{D_{dom}}).
\end{aligned}$$

We note that, if UDA is of interest, an additional classifier (as discussed in Section 1.3.2) will be added with the loss $\mathcal{L}_{cla}$. Thus, the gradient of θ_{E_c} is derived as follows:

$$\theta_{E_c} \xleftarrow{+} -\Delta_{\theta_{E_c}}(\mathcal{L}_{rec} + \mathcal{L}_{adv}^{E_c} + \mathcal{L}_{con} + \mathcal{L}_{sty} + \mathcal{L}_{cla}). \qquad (1.13)$$

Once the training is complete, our model can be applied to image translation in the following ways:

1) For an input image, we utilize the content encoder E_c to extract its content

feature. By conditioning on a randomly sampled domain-specific feature with a selected domain code l_i, generator G would manipulate and output the image in the domain of interest.
2) Given two images of interest, we extract the content feature z_i^c from one image, and the domain-specific feature z_j^d from another (together with its domain code l_j). This can be viewed as example-guided image translation.

It is worth noting that, our disentangled representations are achieved by jointly minimizing domain confusion loss ($\mathcal{L}_{adv}^{D_c}$, $\mathcal{L}_{adv}^{E_c}$), reconstruction loss $\mathcal{L}_{rec}$, content consistency loss $\mathcal{L}_{con}$, and style regression loss $\mathcal{L}_{sty}$. Specifically, we explicitly derive the domain-invariant content feature from input images via domain confusion loss ($\mathcal{L}_{adv}^{D_c}$, $\mathcal{L}_{adv}^{E_c}$) in an adversarial manner, allowing our content encoder E_c to extract domain-invariant features. Moreover, we have the content and style consistency losses ($\mathcal{L}_{con}$ and $\mathcal{L}_{sty}$) deployed in our architecture; the former ensures that the input and the translated images preserve the same content feature representation, while the latter enforces the transformed output to be of the style of interest. Finally, the reconstruction loss $\mathcal{L}_{rec}$ is applied to jointly observe the aforementioned disentangled representation with data recovery guarantees. In Table 1.3, we have ablation studies to support the design of our proposed network in performing representation disentanglement.

Also, as shown in Fig. 1.4, 1.6, 1.9, we show that our proposed model is able to derive disentangled representations from input images of the *seen* domains and producing diverse outputs in the *seen* domain of interest during inference time. We note that, existing state-of-the-art image translation models via representation disentanglement (e.g., UNIT [13], E-CDRD [22], MUNIT [24], DRIT [25], UFDN [1]) cannot generalize to images in *unseen* domains. This also verifies that our model exhibits excellent abilities in decoupling content and style-dependent features for image translation.

doi:10.6342/NTU202301924

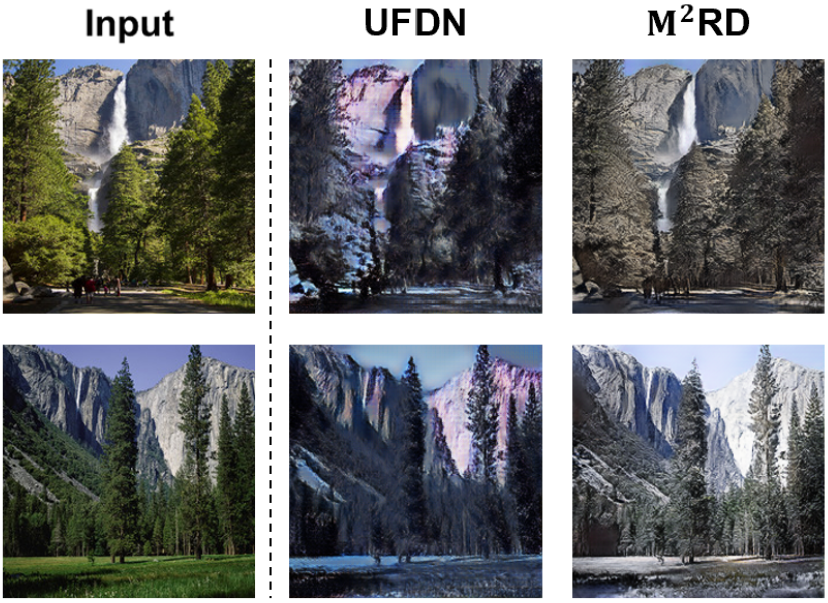


Figure 1.5: Example results of the comparison with UFDN [1] in summer-to-winter translation. Note that since UFDN [1] does not observe and exhibit intra-domain image variety, its output might be irrational in terms of appearance or lighting (e.g., mix of daytime and nighttime appearance), while ours are more realistic and have a higher visual quality.

1.3.5 Comparisons to Recent Models

It is correct that, while our model is related to a recent multi-domain image translation method of UFDN [1], and a number of network modules are shared by this work and ours, multi-modality is the major highlight of our work, plus the introduced feature-level consistency to improve the output image quality. As we noted in Table 1.1, our M^{2}RD further exhibits multi-modal property during the translation/synthesis process, which cannot be achieved by UFDN [1]. However, such extension is not trivial. First, our M^{2}RD needs to derive disjoint domain-specific features (z^d) from the domain-invariant features z^c at the output of the domain encoder (E_d). With detailed model and loss designed are described in our work, we then fit such disentangled domain-specific features to Gaussian distribution priors, allowing the learning of multi-modality in image translation.

doi:10.6342/NTU202301924

As discussed, the domain code (l) in our model serves as supervision, which guides our unified generator to synthesize the output image in the domain of interest. In contrast, UFDN [1] can only perform one-to-one image translation without diversity. In Fig. 1.4 and Table 1.2, we present qualitative and quantitative comparisons respectively to confirm the capability of our M^{2}RD to translate images across multiple domains with sufficient diversity.

Table 1.2: Quantitative comparisons for visual realism and diversity with MUNIT, DRIT, UFDN, and our M^{2}RD on summer-to-winter translation.

		MUNIT	DRIT	UFDN	M^{2}RD (Ours)
Realism	User Study (↑)	21.17%	18.17%	19.33%	**41.33%**
	FID (↓)	85.09 ± 0.77	68.44 ± 0.75	87.69 ± 0.70	$\mathbf{57.76} \pm \mathbf{0.23}$
	LPIPS (I2O) (↓)	0.417 ± 0.003	0.385 ± 0.002	0.758 ± 0.002	$\mathbf{0.339} \pm \mathbf{0.003}$
Diversity	LPIPS (O2O) (↑)	$\mathbf{0.225} \pm \mathbf{0.002}$	0.173 ± 0.002	0.040 ± 0.001	0.196 ± 0.003

Second, we consider to exploit both *inter*-domain and *intra*-domain variation during image translation, while UFDN [1] only observes inter-domain variation. As shown and compared in Table 1.2, the lack of the ability in modeling intra-domain diversity would lead to a discernible drop in visual quality. Take Fig. 1.5 for examples, the domain change in seasons would be viewed as *inter*-domain variations, while the day/night lighting, etc. condition changes are modeled as *intra*-domain variations. Without our derivation of domain-specific features z^d, one cannot produce translated image outputs with satisfactory quality, generating winter scenes with irrational or unrealistic lighting conditions (and thus poor user study results, as shown and compared in Table 1.2).

Third, we employ cycle-consistency loss in our model for feature consistency guarantees, while UFDN [1] does not include such constraints and thus suffers from drops in visual quality in performing image translation. To be more precise, we utilize *content* consistency to preserve content information during the generation process, instead of directly applying *pixel-level* consistency as used in DRIT [25]. Throughout our experiments, we observe that adding data recovery constraints over pixel levels would be overly restrictive and limit the diversity of the image outputs. With the above observations and as summarized in Table 1.2, we show that our model achieved higher LPIPS (O2O) score than DRIT [25] did, which supports the effectiveness of our model in preserving content consistency during image translation. With the above remarks, we believe the technical contributions of this work would be sufficiently unique, which makes our work very different from UFDN [1].

1.4 Experiments

1.4.1 Implementation Details

We utilize PyTorch [41] to implement our model and choose ADAM [42] as the optimizer to train our network, with the learning rate, β_1, and β_2 set as 10^{-4}, 0.5, and 0.999, respectively. In our all experiments, we set the hyperparameters as follows: $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_{rec} = 10$, $\lambda_{con} = 1$, $\lambda_{KL} = 10^{-3}$, $\lambda_{sty} = 10$, and $\lambda_{cla} = 1$.

More details about the network architecture for Summer $\leftrightarrow$ Winter and Photo $\leftrightarrow$ Art datasets are described in the following.

For content encoder E_c, we apply convolutional architecture composed of three convolution layers and four residual blocks. For domain encoder E_d, we implement it by utilizing four convolution layers followed by a fully-connected layer. Also, we use four residual blocks, followed by three deconvolution layers to realize generator G. For content discriminator D_c, it consists of five fully-connected layers. For domain discriminator D_{dom}, we utilize the architecture of PatchGANs [9] that contains six convolution layers, and add two convolution layers for outputting real/fake and domain code prediction respectively.

doi:10.6342/NTU202301924

1.4.2 Datasets

We consider four different categories of image datasets, *i.e.,* digit, face, seasons, and art paint, for performance evaluation:

Digits. The image datasets of *MNIST*, *USPS*, and *Street View House Number (SVHN)* are hand-written digit image datasets, which are viewed as images observed in different domains. MNIST contains 60,000/10,000 images for training/testing, and USPS consists of 7,291/2,007 images for training/testing. SVHN is composed of colored digits images with complex backgrounds and contains 73,257 training images, 26,032 testing images, and 531,131 extra images. All images are converted to RGB images with the size of 32 × 32 × 3 pixels for our experiments.

Faces. We consider facial *photo*, *sketch*, and *paint* images as data in different domains. For facial photo images, we consider the *CelebFaces Attributes dataset (CelebA)* [43], which is a large-scale face image dataset including more than $200K$ celebrity photos annotated with 40 facial attributes. Following the settings of [9, 22, 1], we randomly transfer half of the photos to sketch, then convert the remaining photos into paint images.

Summer ↔ Winter. The Summer ↔ Winter dataset [11] contains natural scene images categorized into summer or winter. The size of all images is 256 × 256 × 3 pixels, and the numbers of images are 1273 and 854 for summer and winter, respectively.

Photo ↔ Art. We choose the photo from Yosemite [11] and the *Art* dataset [44] which collected from Wikiart containing 14 different artists. We conduct our experiments on Monet, Van Gogh, and Ukiyo-e, and also resize all images into 256 × 256 × 3 pixels.

It is worth noting that, while image data across multiple domains are presented during the training stage, we do not observe any cross-domain image pairs when learning our proposed model. This is different from recent translation models like [9, 27] with such requirements.

doi:10.6342/NTU202301924

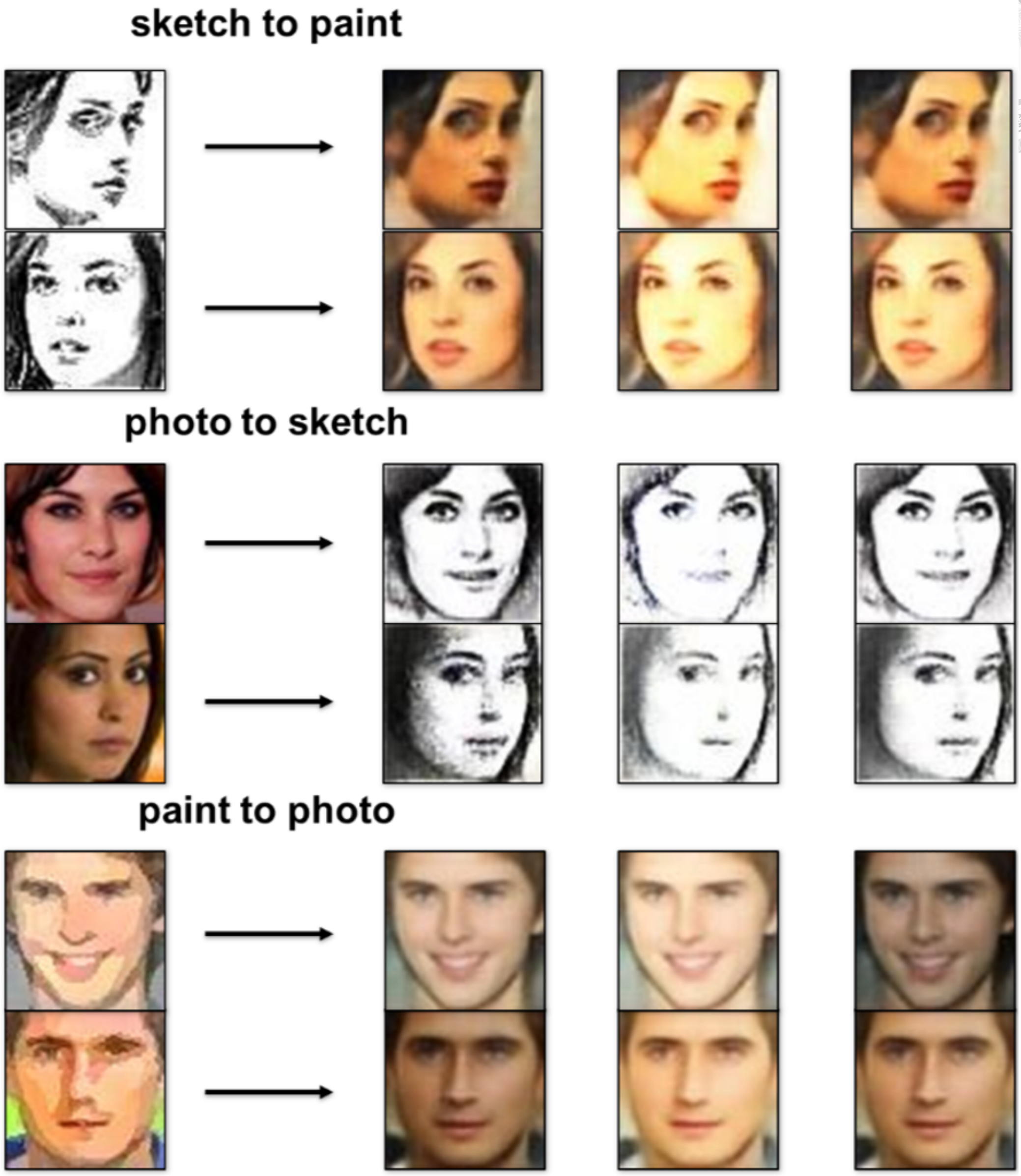


Figure 1.6: Example results of multi-modal image translation for *face* images across multiple domains.

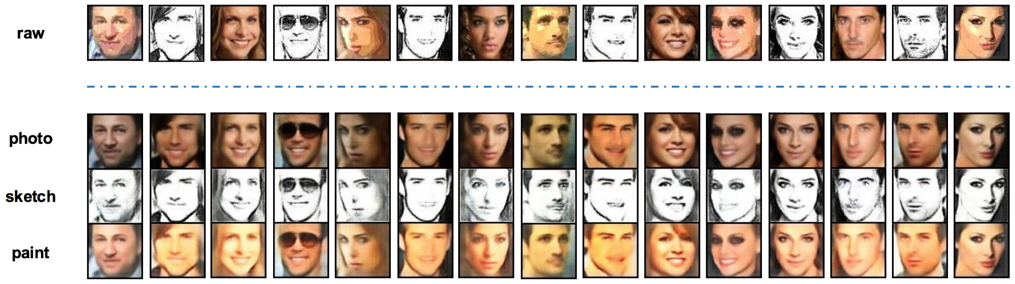


Figure 1.7: Example results of image translation across multiple domains among *photo/sketch/paint*.

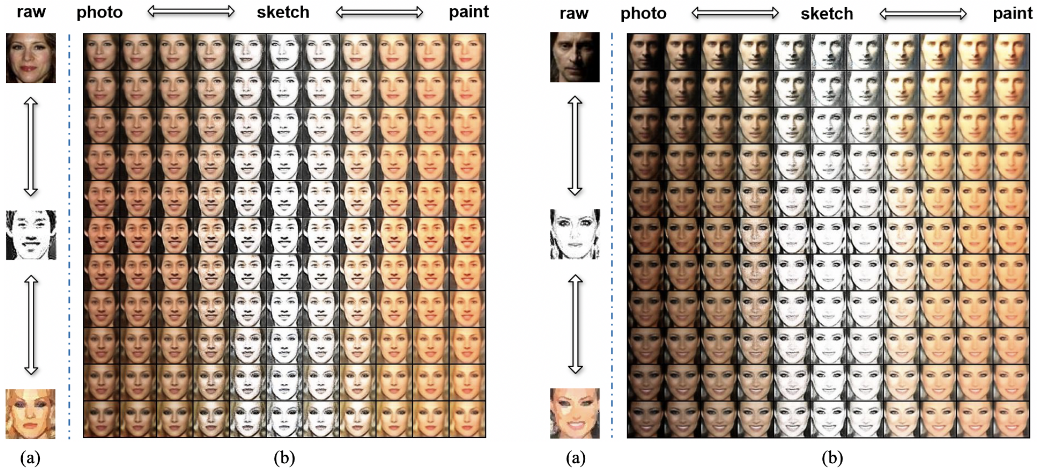


Figure 1.8: Example results of our multi-domain image translations and manipulations. (a) Selected images from three different domains. (b) The horizontal axis shows the cross-domain style interpolation for facial *photo/sketch/paint*, while the vertical axis verifies that the domain-invariant content feature space is continuous.

1.4.3 Multi-domain and Multi-modal Image Translation and Manipulation

Multi-modal image manipulation

In order to provide diversity in the produced image outputs, we first manipulate the latent feature space by sampling the domain-specific feature from a prior Gaussian distribution, concatenated by a desired one-hot domain code. Example results

doi:10.6342/NTU202301924

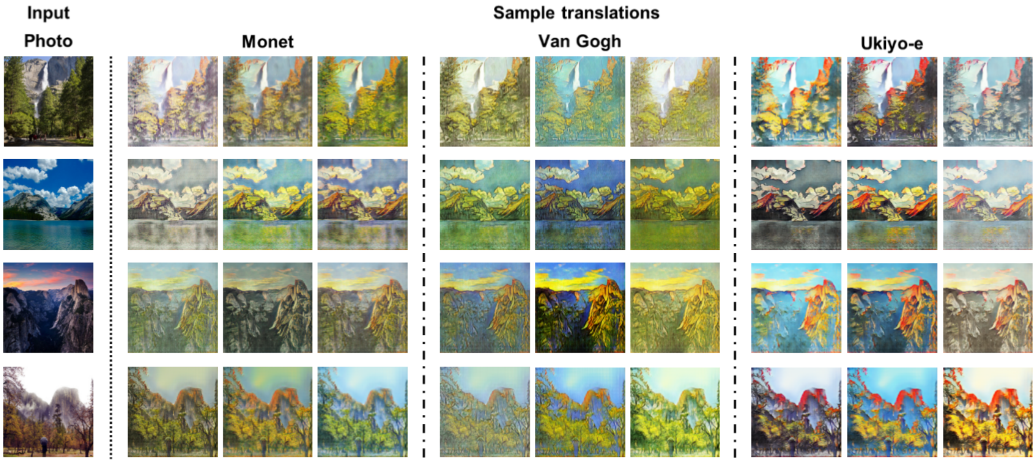


Figure 1.9: Example results of our multi-domain and multi-modal image translations. We translate the input photos into another *Paint style* by manipulating the domain code. Further, by randomly sampling distinct noise vectors, we are able to synthesize output images in the domain of interest with multi-modality.

are shown in Fig. 1.4 on *summer* ↔ *winter* dataset and Fig. 1.6 on *face* dataset, in which multiple outputs in each domain can be produced based on the same input image. Specifically, in Fig. 1.4, we compare our M^{2}RD with the state-of-the-art image-to-image translation methods, showing that our M^{2}RD is capable of synthesizing high-quality output images with diversity. We observe that only injecting noise vectors to the generator of CycleGAN [11], which originally focuses on one-to-one image translation, cannot produce diverse outputs. While UFDN [1] translates images across multiple domains, the generated images mainly belong to one mode and fail to synthesize multi-modal images. Comparing with DRIT [25] and MUNIT [24], DRIT also generates plausible results, and MUNIT produces images with unrealistic style. We also demonstrate that our model without content discriminator (D_c) cannot preserve domain-invariant information well, causing unrealistic and ill-quality results. From the above experiments, the use of our proposed M^{2}RD for multi-modal image translation can be successfully verified.

In addition to qualitative results and comparisons, we further provide additional

doi:10.6342/NTU202301924

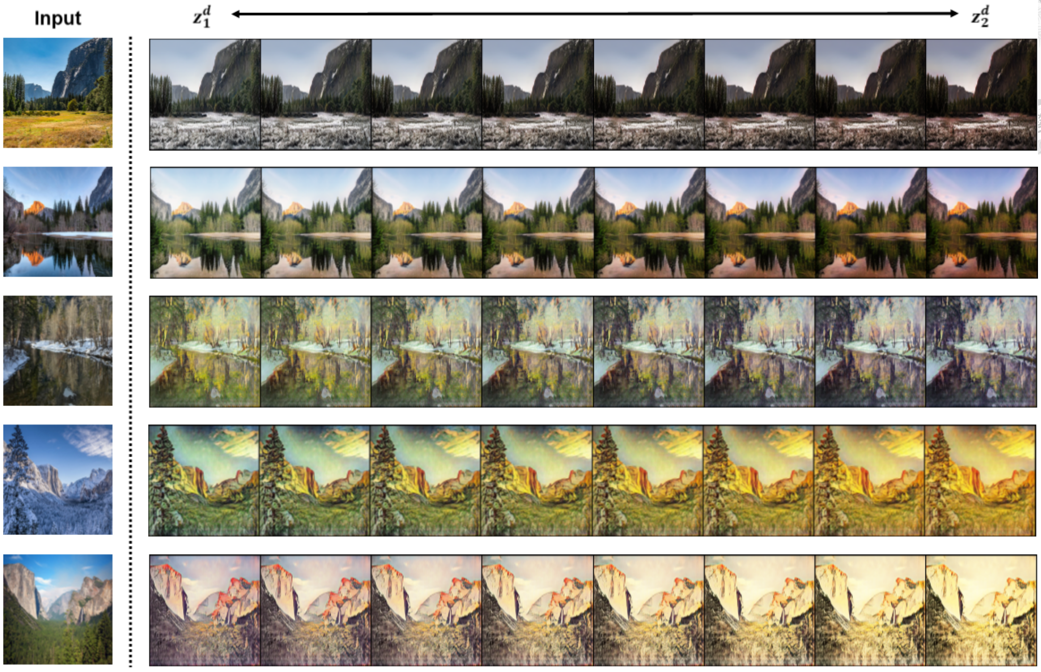


Figure 1.10: Example results of linear interpolation between two sampled random vectors both on *Simmer* ↔ *Winter* and *Photo* ↔ *Art* dataset.

Table 1.3: Ablation studies on summer to winter translation.

		w/o D_c	w/o D_{dom}	w/o $\mathcal{L}_{con}$	w/o $\mathcal{L}_{sty}$	w/o $\mathcal{L}_{KL}$	M^{2}RD (Ours)
Realism	FID (↓)	60.35 ± 0.56	444.48 ± 3.71	73.76 ± 0.97	68.65 ± 0.82	99.24 ± 1.37	**57.76** ± **0.23**
	LPIPS (I2O) (↓)	0.354 ± 0.002	0.976 ± 0.003	0.364 ± 0.002	0.347 ± 0.002	0.397 ± 0.003	**0.339** ± **0.003**
Diversity	LPIPS (O2O) (↑)	0.136 ± 0.002	0.067 ± 0.004	0.187 ± 0.002	0.107 ± 0.001	0.158 ± 0.002	**0.196** ± **0.003**

quantitative comparisons with MUNIT [24], DRIT [25], and UFDN [1], which are known as the state-of-the-art models on image translation.

To assess the visual quality and realism of the synthesized images, we adopt *Frechet Inception Distance* (FID) [45] and *Learned Perceptual Image Patch Similarity* (LPIPS) [46] as the metrics for quantitative evaluation. We compute FID to measure the distance between the generated distribution and the real image input, and we also calculate average Input-to-Output LPIPS, denoted as LPIPS (I2O), to measure the distance between the input image and its corresponding translated outputs (note that lower scores indicate outputs with better visual quality). In addition, we conduct studies by asking 30 users with diverse backgrounds and knowledge

doi:10.6342/NTU202301924

with 20 questions, each contains a given input image and four translated images generated by the above models (including ours), and the user is asked to select the one which he/she feels to be most appropriate/realistic. In Table 1.2, we show that our M^{2}RD outperformed the aforementioned state-of-the-art multi-modal or multi-domain image translation models in all categories. With this experiment, we confirm that our model is capable of producing output images with satisfactory visual quality.

In addition to visual realism, we provide quantitative comparisons for visual diversity by calculating average Output-to-Output LPIPS, denoted as LPIPS (O2O), to measure the distance between the outputs translated from the same input image (note that larger distance values represent output images with more diversity). As shown in Table 1.2, we see that despite UFDN [1] is capable of translating images across multiple domains, it cannot achieve *multi-modal* image translation (with the lowest LPIPS score). More importantly, our model was shown to perform favorably against DRIT [25] and MUNIT [24], which support the ability of our model in synthesizing plausible outputs with sufficient multi-modal diversity. With the above quantitative comparisons, the robustness and superiority of our model can be successfully verified.

Multi-domain image manipulation

We demonstrate the ability of our model to realize image translation across multiple domains using *face* dataset. Given images from an arbitrary domain (i.e., top row in Fig. 1.7), we extract their domain-invariant and domain-specific features, respectively. For translation purposes, we assign and concatenate the above features with different domain codes of interest (e.g., [1, 0, 0] for *photo*, [0, 1, 0] for *sketch*, and [0, 0, 1] for *paint*) for image reconstruction. The translated results were shown in each corresponding column in the bottom row of Fig. 1.7.

Then, given images from different domains (i.e., *photo*, *sketch*, and *paint* in Fig. 1.8a), we extract their domain-invariant (content) features and domain-

doi:10.6342/NTU202301924

specific (style) features. Then, we perform feature interpolation within the same feature type. Using the resulting content/style features with an interpolated domain code, we are able to produce cross-domain image translation outputs. As shown in Fig. 1.8b, outputs in vertical and horizontal axes represent image variants in (domain-invariant) content and (domain-specific) style with the associated domain code, respectively. Observing the diagonal entries of Fig. 1.8b, which shows the extreme translation case, and fully exhibits the effectiveness and robustness in the derived feature representations for multi-domain image manipulation.

In addition to faces, we also demonstrate the use of our model for manipulating hand-written digit images. As shown in Fig. 1.11a and b, by manipulating the domain-specific feature with the desirable domain code (e.g., [1, 0] for USPS/SVHN, and [0, 1] for MNIST), our model is able to convert the USPS and SVHN images into MNIST ones. The above experiments of the use of our proposed M^{2}RD for multi-domain image manipulation are supportive.

Multi-modal translation across multiple domains

As shown in Fig. 1.9, we conduct the experiment of multi-modal image translation across multiple domains on *Photo* ↔ *Art* dataset. By manipulating the domain code (e.g., [0, 0, 0, 1] for *Photo*, [0, 0, 1, 0] for *Monet*, [0 , 1, 0, 0] for *Van Gogh*, and [1, 0, 0, 0] for *Ukiyo-e*), our M^{2}RD is capable of translating given images to the domain of interest. We show that our model successfully captures different *Painting style* and presents clearly distinct results. Furthermore, by randomly sampling different noise vectors from the prior Gaussian distribution, we are able to model the intra-domain variation and perform multi-modal diversity.

For further evaluate the domain-specific (style) latent space derived by M^{2}RD, we perform linear interpolation between two sampled style features as shown in Fig. 1.10. The corresponding results both on *Summer* ↔ *Winter* and *Photo* ↔ *Art* dataset change smoothly and continuously along with the variations of style latent feature.

doi:10.6342/NTU202301924

Table 1.4: A classification accuracy (%) for target domain images. For example, USPS → MNIST denotes USPS and MNIST as source and target domain images, respectively.

	MNIST → USPS	USPS → MNIST	SVHN → MNIST
DANN [14]	-	-	73.85
Associative DA [17]	-	-	93.71
DSN [16]	-	-	82.70
DTN [10]	-	-	84.88
PixelDA [47]	-	95.9	-
DRCN [48]	91.80	73.70	82.00
CoGAN [12]	95.65	93.15	-
ADDA [49]	89.40	90.10	76.00
UNIT [13]	95.97	93.58	90.53
CyCADA [18]	-	-	90.08
ADGAN [50]	92.80	90.80	92.40
CDRD [22]	95.05	94.35	-
SBADA-GAN [51]	97.6	95.0	76.1
UFDN [1]	97.13	93.77	**95.01**
M^{2}RD (Ours)	**98.54**	**98.49**	94.03

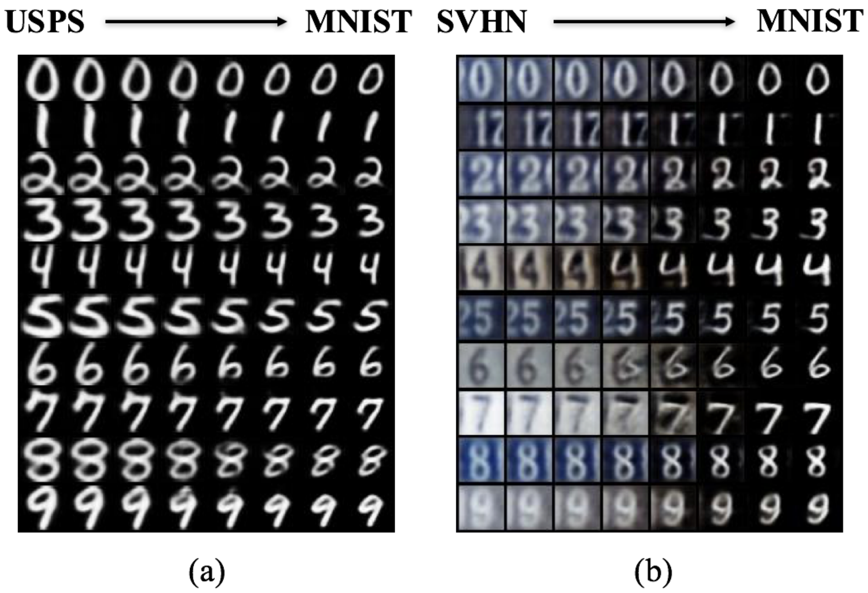


Figure 1.11: Cross-domain continuous image manipulation for (a) USPS → MNIST and (b) SVHN → MNIST.

Quantitative Ablation Study

In addition to the qualitative ablation study (i.e., Fig. 1.4) which partially performs such ablation studies (i.e., our model with and without D_c), we now present additional quantitative ablation studies in Table 1.3 to verify the technical contributions of our work.

As shown in Table 1.3, our model surpassed others in terms of all metrics of FID and LPIPS scores, which confirms the visual quality and diversity achieved by the full model of our M^{2}RD. We observe that, without content discriminator D_c, all scores became inferior since the derived features from content encoder E_c will not be domain-invariant and would carry the domain-specific information, even with the presence of domain-specific feature z^d and domain code l. This supports our network/loss designs for representation disentanglement. Moreover, without domain discriminator D_{dom}, all scores were degenerated significantly due to image details of the outputs across different domains cannot be properly preserved. Next, when the content consistency loss $\mathcal{L}_{con}$ was disabled, the content information would not be preserved well, resulting in poor visual quality and inferior FID/LPIPS scores. If the style regression loss $\mathcal{L}_{sty}$ was removed, we were not able to ensure the style information could be contained well in domain-specific features z^d, and thus lead to lower LPIPS (O2O) scores (i.e., images with poor diversity). Without $\mathcal{L}_{KL}$, we were not able to enforce the encoded domain-specific features to fit the prior Gaussian distribution, and thus failed to exhibit the multi-modal ability in cross-domain image translation. As a result, all the scores based on image realism and diversity dropped drastically. With the above quantitative ablation studies, we confirm the effectiveness and robustness of our M^{2}RD in performing multi-modal image translation across multiple domains.

1.4.4 Unsupervised Domain Adaptation

Finally, we apply our model for cross-domain classification. More specifically, we consider the challenging task of unsupervised domain adaptation (UDA), which aims at classifying images in the target domain while the labels are only available in the source domain during training. We conduct the UDA experiments using the handwritten digit datasets. For instance, MNIST → USPS indicates the use of MNIST as source-domain labeled data, while USPS is in the target domain without any categorical information. As mentioned in Section 1.3.2, UDA can be achieved by our model by adding an extra classifier to recognize the disentangled content features. This classifier is jointly trained with our M^{2}RD.

Table 1.4 compares the results of our model with recent translation-based UDA approaches. For MNIST → USPS, we achieved improved performances over the state-of-the-art methods, and our model performed favorably against others in USPS → MNIST. As for SVHN → MNIST, which is considered to be a more difficult scenario due to significant differences in background, stroke, and illumination, very promising results were reported by our proposed model as well.

In addition to quantitative evaluation, we further provide visualization results to further assess the UDA ability using our derived features. As shown in Fig. 1.12,

doi:10.6342/NTU202301924

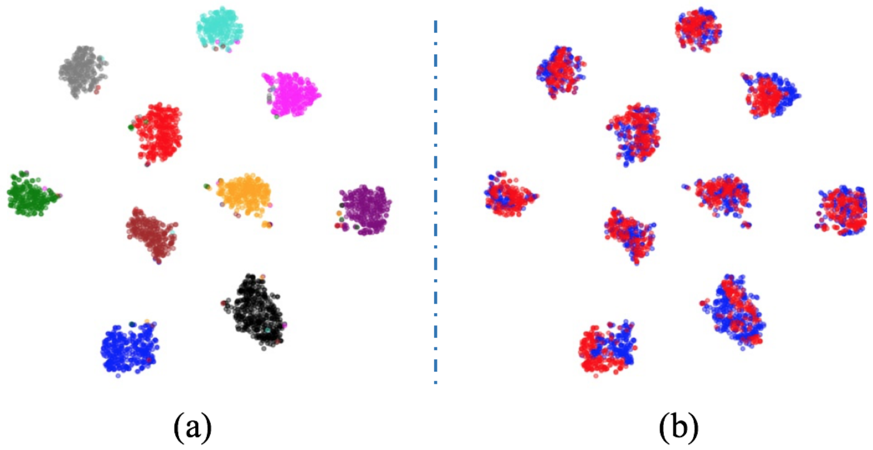


Figure 1.12: t-SNE visualization of the handwritten digit data for USPS → MNIST. Note that different colors indicate data of (a) different digits classes **0-9** and (b) different domains (**source/target**).

we visualize domain-invariant representations of USPS → MNIST using t-SNE. To be more precise, Fig. 1.12a illustrates the image data of 10 categories which were properly separated, while Fig. 1.12b shows the same data associated with different domains (which are close to each other with reduced domain differences).

1.5 Conclusions

In this thesis, we proposed a unified deep learning model of Multi-domain and Multi-modal Representation Disentangler (M^{2}RD). This unique network architecture addresses image manipulation and recognition across multiple domains by properly disentangling feature representation of interest. As a unique characteristic, multi-modal diversity is introduced into our proposed model, which realizes multi-modal image translation during the image manipulation process. In our experiments, we successfully verified that our model produced promising multi-domain and multi-modal image manipulation results using face, seasons, paints, and handwritten digit data, and can be applied to solve unsupervised domain adaptation with satisfactory accuracy.

doi:10.6342/NTU202301924

Chapter 2

Knowledge Transfer for Semantic Categories

Zero-shot learning (ZSL) requires one to associate visual and semantic information observed from data of seen classes, so that test data of unseen classes can be recognized based on the described semantic representation. Aiming at synthesizing visual data from the given semantic inputs, hallucination-based ZSL approaches might suffer from mode collapse and biased problems due to the lack of ability in modeling the desirable visual features for unseen categories. In this thesis, we present a generative model of Cross-Modal Consistency GAN (CMC-GAN), which performs semantics-guided intra-category knowledge transfer across image categories, so that data hallucination for unseen classes can be achieved with proper semantics and sufficient visual diversity. In our experiments, we perform standard and generalized ZSL on four benchmark datasets, confirming the effectiveness of our approach over that of state-of-the-art ZSL methods.

2.1 Introduction

Deep learning approaches have shown promising performances in several computer vision tasks like object classification [57, 58], detection [59, 60], and segmenta-

doi:10.6342/NTU202301924

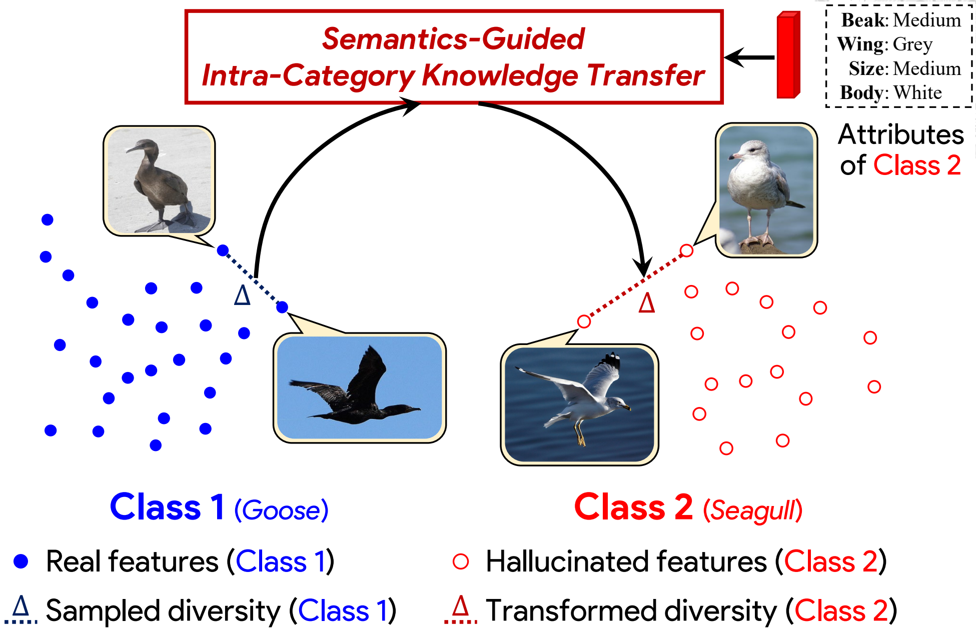

Semantics-Guided
Intra-Category Knowledge Transfer
Beak: Medium
Wing: Grey
Size: Medium
Body: White
Attributes of Class 2
Δ
Δ
Class 1 (Goose)
Class 2 (Seagull)
● Real features (Class 1)
○ Hallucinated features (Class 2)
Δ Sampled diversity (Class 1)
Δ Transformed diversity (Class 2)

Figure 2.1: Illustration of transferring semantics-guided intra-category knowledge for hallucinating visual features of unseen class with proper semantics and visual appearance. Note that visual diversities across seen and unseen classes are learned and preserved.

Table 2.1: Comparisons with recent approaches on zero-shot learning.

	Cross-Modal Embedding based ZSL			Data Generation based ZSL			
	SJE [52]	DEM [53]	CADA-VAE [54]	f-CLSWGAN [2]	f-VAEGAN [55]	LsrGAN [56]	Ours
Cross-modal association	✓	✓	✓	✓	✓	✓	✓
Hubness alleviation	-	✓	✓	✓	✓	✓	✓
Pseudo visual data synthesis	-	-	-	✓	✓	✓	✓
Mode-collapse alleviation	-	-	-	-	✓	-	✓
Modeling intra-class diversity	-	-	-	-	-	-	✓

tion [61, 62]. Deep neural networks utilized in such applications are typically trained in a fully supervised fashion, which requires a large amount of labeled data for each category of interest. However, real-world data often follow the long-tailed distribution, and thus collecting a sufficient amount of annotated training samples would not be applicable to every class. When it comes to applications like the recognition of rare species or medical image analysis, only a few or even no visual samples are at hand, which leads to a severe overfitting problem. Hence, the scalability and applicability of existing supervised deep learning models would be limited.

To generalize the model learned from seen classes to handle unseen class data, a learning paradigm called *zero-shot learning* (ZSL) [63, 64] is proposed to transfer the knowledge across such categories. More precisely, ZSL aims at recognizing instances of *unseen* classes without observing any visual cues during training. With the aid of side information (*e.g.,* image attributes or texture descriptions), ZSL approaches focus on relating visual and semantic domains, so that data presented in unseen categories can be processed and classified accordingly. For a more challenging yet practical scenario, *generalized zero-shot learning* (GZSL) [65, 66] requires one to not only recognize data of previously unseen classes during inference, but also exhibit the ability to classify images of seen categories as well.

One group of methods formulate ZSL as a visual-semantic matching task in a *deterministic* manner, and seeks to embed both visual representations and the corresponding class attributes into a shared feature space for classification purposes. With such feature representations derived, matching query images of unseen categories and their semantic representation can be performed. For example, methods like [52, 67, 68] choose to project visual data of seen classes into a space spanned by all class attributes. However, there is no guarantee to preserve both inter-class divergence and intra-class variation after embedding high-dimensional visual data into a less informative low-dimensional class attribute space. This might lead to the *hubness problem* [69, 53, 70], making the projected data easily clustered as hubs and thus hamper the recognition performance. In order to alleviate this problem, [69, 53, 70, 71] learn the mapping from attribute space to visual space instead. However, directly mapping semantic attributes to visual space might lead to ambiguity during ZSL classification, since the separation between objects with overlapping attributes might not be easily achieved.

In contrast to cross-modal matching, another group of approaches utilizes generative models to facilitate the learning of embedding spaces for ZSL. Inspired

doi:10.6342/NTU202301924

by generative adversarial networks (GANs) [23], methods like [2, 72, 55, 73] learn a feature generator that hallucinates visual data from class-level attributes, so that classifiers can be trained accordingly. While promising results were reported, the generation process of existing GAN-based methods is generally based on a randomly sampled noise vector conditioned on the class/attribute information. In other words, intra-class diversity simply relies on seen-class data sampled during GAN training. This might suffer from *mode collapse* [74] and *biased problems* [65], resulting in synthesized data visually similar to each other or fitting particular seen categories, respectively.

In this thesis, we aim at exploiting and transferring inherent intra-class visual variation across image categories, so that hallucinating unseen class data can be realized, as illustrated in Fig. 2.1. In particular, we propose a novel *Cross-Modal Consistency GAN (CMC-GAN)* for ZSL hallucination, as depicted in Fig. 2.2. Different from existing GAN-based models for data hallucination, CMC-GAN jointly observes semantic and visual data as inputs, with characteristics described as follows: (1) **Unpaired visual and semantic training data**: When learning to synthesize visual data, CMC-GAN jointly takes class-specific attributes and a pair of visual data as the inputs, which are not required to be sampled from the same category during training. This allows us to generate visual data across image categories. (2) **Diversity preserving across categories**: Without requiring visual and semantic correspondences during training, we introduce a unique module of *semantics-guided intra-category knowledge transfer*, which translates the observed intra-class variation from one class to another with the guidance of the semantic attributes of that class as illustrate in Fig. 2.1. This allows us to transfer and synthesize visual data for unseen classes. (3) **Data generation with semantics and visual diversity preservation**: Our CMC-GAN is able to either synthesize visual data given only semantic information, or to produce such data with additional visual diversity observed from other seen classes. By observing semantic and visual diversity consistency at different modules/outputs, our CMC-GAN is capable

doi:10.6342/NTU202301924

of hallucinating desirable outputs across (seen or unseen) image categories. In Table 2.1 (and as discussed in Sec. 2.2), we compare several ZSL methods, and highlight the differences between ours and such models.

The contributions of this thesis are summarized below:

- We present a Cross-Modal Consistency GAN (CMC-GAN) model for ZSL, allowing hallucination of visual data while preserving and manipulating desirable semantic and visual diversity information.

- Instead of proposing *novel* network designs or loss functions, our framework focuses on translating visual diversity into categories of interest, so that hallucination of unseen class data can be achieved.

- We uniquely observe semantic and visual diversity consistency at attribute and feature levels, so that mode collapse or biased problems can be alleviated.

- Our CMC-GAN tackles both ZSL and GZSL tasks using benchmark datasets, and performs favorably against state-of-the-art embedding and hallucination-based approaches.

2.2 Related Works

2.2.1 Cross-Modal Embedding

Due to the lack of visual data of unseen classes during training, one cannot directly train visual classifiers to recognize such data. Existing ZSL methods typically adopt semantic side information for bridging the gap across seen and unseen classes, with a feature space shared across visual and semantic modalities for classification purposes. Several approaches have been proposed to learn visual-semantic embedding for matching visual and semantic information from training data of seen classes. For example, [75, 52, 76, 77, 67, 68] map the visual features into class attribute space and predict the labels by finding the most matching

doi:10.6342/NTU202301924

class attribute. In contrast of the above methods, [69, 53, 70] turn to approach ZSL by projecting class semantic attribute to visual space. While boosting the recognition accuracy, the mapping is learned in a deterministic fashion, which raises the ambiguity that one class attribute would correspond to several possible images. Recently, CADA-VAE [54] modifies variational auto-encoder (VAE) [34] for aligning the data from different modalities into a shared latent space in a probabilistic way.

Very recently, CLIP [78] is a cross-modal pre-training approach, which allows zeros-shot transfer to unseen datasets. They adopt contrastive learning to maximize the similarity of the corresponding image-text pair, while repulsing all other texts from the anchor image. Once the pre-training is complete, CLIP [78] is capable of using the trained text encoder to transform the semantic attributes/class labels to the class prototypes, which can be viewed as the weights of the classifier and thus enables zero-shot learning. However, due to the inherent heterogeneity gap across visual and semantic modalities, plus the unbalance nature between instance-wise visual features and class-wise semantic attributes, how effectively learning a latent space to relate such cross-modality information would still be a challenging task. Instead of only considering the alignment across different modalities, our CMC-GAN aims at exploiting the intra-class variation and synthesizing the visual features based on the associated semantic attributes with such derived intra-class diversity.

2.2.2 Data Generation

Zero-shot learning. A number of ZSL approaches [2, 72, 55, 73, 79, 80, 56] extend Generative Adversarial Networks (GANs) [23] to generate pseudo training data from their semantic attributes due to the fact that the conditional GAN generator is capable of synthesizing more discriminative visual features than other types of generative models (*e.g.,* conditional VAEs [81]). With pseudo visual data synthesized, one can train a standard KNN classifier directly for recognizing test

doi:10.6342/NTU202301924

data of both seen and unseen classes (*i.e.,* GZSL). f-CLSWGAN [2] is proposed to apply WGAN [82] as the feature generator to synthesize visual samples, conditioned on class attributes with random noise which depicts intra-class variation. Most recently, Cycle-WGAN [72] and DASCN [80] augment the f-CLSWGAN framework by adding the cycle consistency for the semantic attribute using L_2 loss and dual-adversarial objective respectively. This is to enforce the generated visual features to fully represent their corresponding semantic attributes. While Cycle-WGAN [72] and DASCN [80] apply cycle consistency to enforce the *semantic attributes* to be properly exploited, they do *not* observe or preserve *intra-category diversity* consistency for ensuring sufficient visual diversity during visual feature hallucination. In addition, [55] incorporate VAE [34] with [2] to facilitate the capability of capturing real visual distribution. The recent state-of-the-art Lsr-GAN [56] leverages the semantic relationships between seen and unseen classes to encourage the produced visual features to preserve the relationships observed from semantic space. Though the above models introduce the capability of generating visual features which preserve their corresponding semantic meanings, they are generally not designed to observe or preserve proper intra-class variation during the generation process, which would be a critical factor for approximating the data distributions of distinct classes.

Few-shot learning. In contrast to ZSL, another learning paradigm called few-shot learning (FSL) aims at recognizing novel classes with very few training data. In FSL, data hallucination is also considered as a common technique. For instance, [83] transfers the analogy relation from a pair of images from a known class to images of a novel class. However, they require ad-hoc techniques to select base image categories and their data pairs for performing the above transfer and hallucination process. [84] and [85] capture intra-class deformations from different paired images sampled from the same class for perturbing samples in the support set. While [83], [84], and [85] share the same goal (as ours does) to derive the

doi:10.6342/NTU202301924

inherent intra-class variation from desirable classes, they cannot directly apply for ZSL since their methods are *not* designed to handle data across different modalities. Moreover, the ability to generalize such intra-class knowledge from known classes to unseen classes is not guaranteed for FSL approaches. Thus, one cannot directly apply and extend the above models for ZSL. In summary, focusing on ZSL/GZSL tasks, our proposed model is able to exploit class-specific intra-class diversity, and conditionally transform such information for hallucinating visual data of unseen classes.

Visual diversity of GANs. To encourage visual diversity and prevent the mode collapse problems common in conditional GANs, recent works MSGAN [86] and DRIT++ [87] introduce a mode seeking regularization that maximizes the ratio of the distance between synthesized output images with respect to the distance between their corresponding latent vectors. With this regularization, the visual diversity of generated images could be enforced and the mode collapse problem of conditional GANs would be alleviated. Different from the above methods that focus on enlarging the distance of synthesized images for producing diverse outputs, our approach learns to encourage the derived intra-category diversity of the seen categories to be properly preserved when hallucinating visual features of seen and unseen classes.

2.3 Proposed Method

2.3.1 Problem Formulation and Algorithm Overview

We first define the notation to be used in this thesis. For ZSL, the training data are the seen-class data $\mathcal{D}_S = \{(x, a, y) | x \in \mathcal{X}_S, a \in \mathcal{A}, y \in \mathcal{Y}_S\}$, where $\mathcal{X}_S$ is the set of visual features from seen classes, $\mathcal{A}$ and $\mathcal{Y}_S$ denote the associated attribute and label sets, respectively. We note that the image feature x is extracted by a pre-trained convolutional neural network (*e.g.,* ResNet-50 [58]), with the

doi:10.6342/NTU202301924

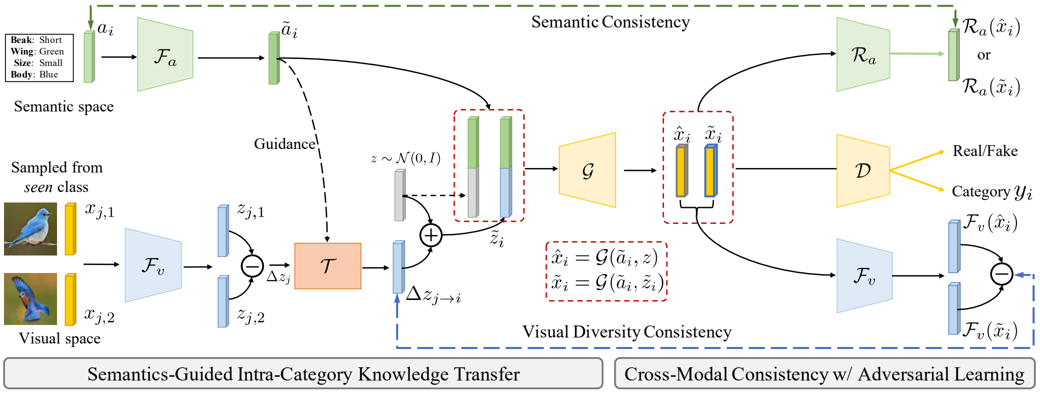

Semantic Consistency
Beak: Short
Wing: Green
Size: Small
Body: Blue
a_i
$\mathcal{F}_a$
$\tilde{a}_i$
Semantic space
$\mathcal{R}_a$
$\mathcal{R}_a(\hat{x}_i)$
or
$\mathcal{R}_a(\tilde{x}_i)$
Guidance
$z \sim \mathcal{N}(0, I)$
Sampled from
seen class
$x_{j,1}$
$x_{j,2}$
Visual space
$\mathcal{F}_v$
$z_{j,1}$
$z_{j,2}$
Δz_j
$\mathcal{T}$
$\Delta z_{j \to i}$
$\tilde{z}_i$
$\mathcal{G}$
$\hat{x}_i$
$\tilde{x}_i$
$\mathcal{D}$
Real/Fake
Category y_i
$\hat{x}_i = \mathcal{G}(\tilde{a}_i, z)$
$\tilde{x}_i = \mathcal{G}(\tilde{a}_i, \tilde{z}_i)$
$\mathcal{F}_v(\hat{x}_i)$
$\mathcal{F}_v(\tilde{x}_i)$
Visual Diversity Consistency
Semantics-Guided Intra-Category Knowledge Transfer
Cross-Modal Consistency w/ Adversarial Learning

Figure 2.2: Architecture of Cross-Modal Consistency GAN (CMC-GAN). The module of Semantics-Guided Intra-Category Knowledge Transfer contains an attribute encoder $\mathcal{F}_a$, a visual encoder $\mathcal{F}_v$, a semantic-conditioned transformation $\mathcal{T}$, and a generator $\mathcal{G}$. The adversarial training includes the discriminator $\mathcal{D}$ with an auxiliary classifier, an attribute regressor $\mathcal{R}_a$ and a shared $\mathcal{F}_v$, enforcing semantic and visual consistencies at attribute/feature levels. Note that a_i and z_j denote the attribute and visual features for classes i and j, respectively. Δz_j and $\Delta z_{j \to i}$ indicate the intra-class variations derived from and translated for the corresponding classes. Note that $\ominus$ and $\oplus$ represent difference and addition mapping functions, respectively; both are realized by a single-layer neural network.

corresponding label y and attribute vector a. As for unseen classes for inference, we only observe $\mathcal{D}_U = \{(a, y) | a \in \mathcal{A}, y \in \mathcal{Y}_U\}$ with $\mathcal{Y}_U$ as the associated label set. Note that $\mathcal{Y}_S \cap \mathcal{Y}_U = \varnothing$.

The task of ZSL is to learn a mapping between $\mathcal{X}$ and $\mathcal{A}$, so that the associated $\mathcal{Y}$ can be determined accordingly. For conventional ZSL, one typically focuses on learning $f_{ZSL} : x \to \mathcal{Y}_U$ for recognizing the input x as one of the unseen classes. As for *generalized ZSL* (as considered in this thesis), one needs to perform a more challenging learning task of $f_{GZSL} : x \to \mathcal{Y}_S \cup \mathcal{Y}_U$, where x can be drawn from either seen or unseen classes.

2.3.2 Cross-Modal Consistency GAN for Data Hallucination

The proposed network of our Cross-Modal Consistency GAN (CMC-GAN) comprises a semantics-guided intra-category knowledge transfer module, followed by a GAN based framework for data hallucination. The former aims at extracting and translating the intra-class diversity from one class to another, while these two classes are not necessarily identical. As for the latter module, it is trained to synthesize visual features given the input semantic feature (*e.g.,* class or attribute embedding), or to produce such data with additional visual diversity observed from other seen classes. Thus, semantic and visual diversity are jointly served as conditions for training our CMC-GAN, while the resulting visual and semantic consistency need to be realized in an adversarial learning scheme. In the following sub-sections, we will detail the designs and properties of our CMC-GAN.

Semantics-Guided Intra-Category Knowledge Transfer Across Image Classes

In order to exploit visual diversity across image categories, we deploy a module of semantics-guided intra-category knowledge transfer, as illustrated in the left-hand side of Fig. 2.2. The class attribute a_i of class i and the visual feature pair $\{x_{j,1}, x_{j,2}\}$ sampled from the same or another seen class j are the inputs to this module. While the class attribute is used to derive the semantic prototype $\tilde{a}_i$ via the attribute encoder $\mathcal{F}_a$, we particularly extract the visual latent features $\{z_{j,1}, z_{j,2}\}$ and calculate its visual difference Δz_j through the visual encoder $\mathcal{F}_v$. To be more specific, the visual difference between the input visual pair is derived by:

$$\Delta z_j = f_{\text{diff}}(z_{j,1}, z_{j,2}), \tag{2.1}$$

where the difference mapping function f_{diff} is realized by a single-layer perceptron (*i.e.,* $\ominus$ in Fig. 2.2). Compared to the simple element-wise subtraction between two visual features, the use of f_{diff} allows us to capture the visual/conceptual variation observed from the input visual feature pair.

In order to translate the above visual diversity observed from class j into class i

of interest, the semantic-conditioned feature transformation block $\mathcal{T}$ is introduced in this module. As depicted in Fig. 2.2, this transformation block takes both the visual difference Δz_j and the semantic prototype $\tilde{a}_i$, aiming at converting the visual diversity across the above categories.

To ensure the produced visual output of the feature transformation block $\mathcal{T}$ would match desirable semantic and visual appearance information, we advance the techniques of recent deep style transfer network like [6, 88, 89]. In these proposed style transfer models, one simply applies instance normalization as feature-wise transformation, and translates the style of a content image to fit the distribution (*i.e.,* mean and standard deviation) of the guided one. Inspired by such techniques, our feature transformation block $\mathcal{T}$ (or $\mathcal{T}_{j\to i}$) takes the semantic prototype of the target class i as guidance, with the mean and standard deviation of semantic prototype $\tilde{a}_i$ as the shifting and scaling parameters, respectively. We then transform the visual diversity Δz_j from class j to align with that of class i. The above process is done by feature normalization using the derived mean and standard deviation from $\tilde{a}_i$, *i.e.,* we calculate $\Delta z_{j\to i}$ by:

$$\Delta z_{j\to i} = \mathcal{T}_{j\to i}(\Delta z_j) = \sigma(\tilde{a}_i) \cdot \frac{\Delta z_j - \mu(\Delta z_j)}{\sigma(\Delta z_j)} + \mu(\tilde{a}_i), \tag{2.2}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ indicate the mean and standard deviation operations, respectively. It is worth repeating that, such normalization-based transformation has shown promising abilities in image translation [6, 88, 89].

With $\Delta z_{j\to i}$ now describing the intra-class information for class i, we can complete the generation of the visual feature $\tilde{x}_i$. We start from a random vector $z \sim \mathcal{N}(0, I)$ sampled from a Gaussian distribution. By introducing a single-layer perceptron f_{add} (*i.e.,* $\oplus$ in Fig. 2.2), we take both z and the transformed offset $\Delta z_{j\to i}$ as the inputs, and produce $\tilde{z}_i$ by $\tilde{z}_i = f_{\text{add}}(\Delta z_{j\to i}, z)$.

As depicted in Fig. 2.2, our *feature generator* $\mathcal{G}$ is conditioned on the semantic prototype $\tilde{a}_i$ and takes either the random noise z or $\tilde{z}_i$ to produce the associated visual features $\hat{x}_i$ or $\tilde{x}_i$ for class i, *i.e.,*

$$\begin{aligned} \hat{x}_i &= \mathcal{G}(\tilde{a}_i, z), \quad \text{where } z \sim \mathcal{N}(0, I) \\ \tilde{x}_i &= \mathcal{G}(\tilde{a}_i, \tilde{z}_i), \quad \text{where } \tilde{z}_i = f_{add}(\Delta z_{j\rightarrow i}, z). \end{aligned} \tag{2.3}$$

Note that $\hat{x}_i$ denotes the attribute-conditional visual feature synthesized from a random noise input, while $\tilde{x}_i$ is the one further exhibiting intra-category diversity translated from class j. With a sufficient amount of $\{\hat{x}_i, \tilde{x}_i\}$ being produced for class i, one can train classifiers to recognize test samples of class i. This is how our CMC-GAN serves as a data hallucination model for ZSL, and the reason why the above process is able to perform semantics-guided intra-category knowledge transfer to produce data for unseen classes.

Cross-Modal Consistencies in Semantics and Visual Diversity

To ensure that the visual diversity of class j would be translated into that of class i, our CMC-GAN requires additional guidance during training in addition to image authenticity. That is, for the generated outputs $\hat{x}_i$ and $\tilde{x}_i$, we need to ensure that their semantic information properly aligns with that of a_i, while their visual diversity would be preserved during the generation process.

In light of the above properties, our CMC-GAN needs to produce visual data with semantic and visual diversity preservation. Firstly, to ensure the outputs $\hat{x}_i$ and $\tilde{x}_i$ sufficiently represent visual data of class i, we observe **semantic consistency** at the attribute level in our network. That is, we apply an *attribute regressor* $\mathcal{R}_a$ to map visual outputs $\hat{x}_i$ and $\tilde{x}_i$ into the semantic attributes $\mathcal{R}_a(\hat{x}_i)$ and $\mathcal{R}_a(\tilde{x}_i)$, and we relate such regressed outputs to the attribute input a_i. In other words, the semantic consistency $\mathcal{L}_{con}^{A}$ is thus defined as follows:

$$\mathcal{L}_{con}^{A} = \frac{1}{2}(||a_i - \mathcal{R}_a(\hat{x}_i)||_2^2 + ||a_i - \mathcal{R}_a(\tilde{x}_i)||_2^2), \tag{2.4}$$

where $a_i \sim \mathcal{A}$ represents the semantic attributes of class i.

On the other hand, we need to enforce that the produced $\{\hat{x}_i, \tilde{x}_i\}$ preserve the visual diversity observed from class j (*i.e.,* $\Delta z_{j\rightarrow i}$), while its attribute information is

doi:10.6342/NTU202301924

aligned with that of class i. This would alleviate possible mode collapse and biased problems as later verified. To ensure this property, our CMC-GAN enforces **visual diversity consistency** at the feature level, which calculates the corresponding loss $\mathcal{L}^{V}_{con}$ as follows:

$$\mathcal{L}^{V}_{con} = ||\Delta z_{j \to i} - f_{diff}(\mathcal{F}_v(\hat{x}_i), \mathcal{F}_v(\tilde{x}_i))||^2_2. \tag{2.5}$$

Recall that the pair of visual features $\{\hat{x}_i, \tilde{x}_i\}$ are synthesized with the condition of semantic prototype $\tilde{a}_i$, based on either random noise z or the one with added intra-class diversity $\tilde{z}_i$, as described in (2.3). As for the difference mapping function f_{diff}, we apply the same single-layer perceptron as described in (2.1).

Therefore, to preserve both semantic and visual diversity information during data hallucination, we have the *cross-modal consistency loss* $\mathcal{L}_{cross}$ calculated as follows:

$$\mathcal{L}_{cross} = \mathcal{L}^{A}_{con} + \mathcal{L}^{V}_{con}. \tag{2.6}$$

Adversarial Learning and Full Objective

As presented above, our proposed CMC-GAN performs semantics-guided intra-category knowledge transfer, which translates intra-class visual diversity observed from particular image pairs from one class to another, while the produced visual features would describe desirable semantic and visual information. To train our CMC-GAN, we follow a number of techniques that are widely applied in GAN and image translation works, as we now discuss.

Firstly, we train the generator $\mathcal{G}$ in CMC-GAN in an adversarial manner by utilizing the discriminator $\mathcal{D}$, which is learned to determine not only the authenticity of the observed visual features, but also their categorical information. In other words, we require the discriminator to predict both the correct realness and the desirable class labels. Following [2], our model is trained on seen data using the

doi:10.6342/NTU202301924

Algorithm 1: Cross-Modal Consistency Generative Adversarial Network (CMC-GAN)

```
Input: Attribute encoder $\mathcal{F}_a$, visual encoder $\mathcal{F}_v$, semantic-conditioned
       transformation $\mathcal{T}$, generator $\mathcal{G}$, discriminator $\mathcal{D}$, attribute regressor
       $\mathcal{R}_a$, $f_{diff}$, and $f_{add}$
Data: Seen-class data $\mathcal{D}_S = \{(x, a, y) | x \in \mathcal{X}_S, a \in \mathcal{A}, y \in \mathcal{Y}_S\}$
Output: $\mathcal{F}_a$, $\mathcal{F}_v$, $\mathcal{T}$, $\mathcal{G}$, $f_{diff}$, and $f_{add}$
for Iters. of whole model do
    Randomly sample a minibatch $\{(x_i, a_i, y_i), (x_{j,1}, x_{j,2})\}$ from $D_S$ ;
    Semantic-Guided Intra-Category Knowledge Transfer
    $\tilde{a}_i = \mathcal{F}_a(a_i)$;
    $z_i = \mathcal{F}_v(x_i), z_{j,1} = \mathcal{F}_v(x_{j,1}), z_{j,2} = \mathcal{F}_v(x_{j,2})$;
    $\Delta z_j = f_{\text{diff}}(z_{j,1}, z_{j,2})$ (Eq. 1);
    $\Delta z_{j \to i} = \mathcal{T}_{j \to i}(\Delta z_j) = \sigma(\tilde{a}_i) \cdot \frac{\Delta z_j - \mu(\Delta z_j)}{\sigma(\Delta z_j)} + \mu(\tilde{a}_i)$ (Eq. 2);
    $z \sim \mathcal{N}(0, I)$;
    $\tilde{z}_i = f_{add}(\Delta z_{j \to i}, z)$;
    $\hat{x}_i = \mathcal{G}(\tilde{a}_i, z)$, $\tilde{x}_i = \mathcal{G}(\tilde{a}_i, \tilde{z}_i)$ (Eq. 3);
    Cross-Modal Consistency with Adversarial Learning
    Compute $\mathcal{L}_{con}^{A}$ (Eq.4) and $\mathcal{L}_{con}^{V}$ (Eq. 5);
    Compute $\mathcal{L}_{WGAN}$ (Eq.7) and $\mathcal{L}_{CLS}$ (Eq. 8);
    for Iters. of updating $\mathcal{F}_v$, $\mathcal{F}_a$, $\mathcal{R}_a$, $\mathcal{G}$, $f_{diff}$, $f_{add}$ do
        $\theta_{\{\mathcal{F}_v, \mathcal{F}_a, \mathcal{R}_a, \mathcal{G}, f_{diff}, f_{add}\}} \overset{+}{\leftarrow}$
          $-\Delta_{\{\mathcal{F}_v, \mathcal{F}_a, \mathcal{R}_a, \mathcal{G}, f_{diff}, f_{add}\}}(\mathcal{L}_{con}^{A} + \mathcal{L}_{con}^{V} - \mathcal{L}_{WGAN} + \mathcal{L}_{CLS})$
    end
    for Iters. of updating $\mathcal{D}$ do
        $\theta_{\mathcal{D}} \overset{+}{\leftarrow} -\Delta_{\theta_{\mathcal{D}}}(\mathcal{L}_{WGAN})$
    end
end
```

doi:10.6342/NTU202301924

objective of WGAN [82] with an auxiliary classification loss:

$$\begin{aligned}\mathcal{L}_{WGAN} &= \mathbb{E}_{(x,a)}[\mathcal{D}(x,a)] - \mathbb{E}_{(x_g,a)}[\mathcal{D}(x_g,a)] \\ &\quad - \lambda \mathbb{E}_{(\bar{x},a)}[(\|\nabla_{\bar{x}}\mathcal{D}(\bar{x},a)\|_2 - 1)^2],\end{aligned} \tag{2.7}$$

$$\mathcal{L}_{CLS} = -\mathbb{E}_{(x_g,y)}[\log P(y|x_g)], \tag{2.8}$$

where $x_g \in \{\hat{x}, \tilde{x}\}$ represents the synthetic samples generated by $\mathcal{G}$, $\bar{x} \sim \alpha x + (1-\alpha)x_g$ with $\alpha \sim U(0,1)$ sampled from a uniform distribution, and λ is the penalty coefficient. Also, $P(y|\tilde{x})$ denotes the predicted conditional probability that sample x_g belongs to its true class label y.

In summary, the total loss function $\mathcal{L}$ sums up losses defined in (2.6), (2.7), and (2.8), learning modules of $\mathcal{F}_v$, $\mathcal{F}_a$, $\mathcal{R}_a$, $\mathcal{G}$, f_{diff}, f_{add}. The training and implementation details are presented in Section 2.4.1. The pseudo code of the proposed CMC-GAN is summarized in Algorithm 1.

2.3.3 (Generalized) Zero-Shot Recognition

As depicted in Fig. 2.3, once the training is complete, we utilize the CMC-GAN to synthesize visual feature data conditioned on the semantic embedding $\tilde{a}$ of unseen classes, together with either randomly sampled noise vector $z \sim \mathcal{N}(0, I)$ or that with intra-class visual diversity transferred from the seen classes. With a sufficient amount of visual data produced for unseen classes (as referred in Fig. 2.8), a standard k-nearest neighbor (KNN) classifier $\mathcal{C}$ is trained to recognize test data of unseen (and seen) classes. More specifically, given a test input feature x, the class label y^* which has the maximum softmax score produced by this classifier is selected as the classification result, *i.e.,*

$$y^* = \arg\max_{y \in \mathcal{Y}} p(y|x; \mathcal{C}), \tag{2.9}$$

where $\mathcal{Y} = \mathcal{Y}_U$ for ZSL, and $\mathcal{Y} = \mathcal{Y}_S \cup \mathcal{Y}_U$ for GZSL settings, respectively. Algorithm 2 summarizes our process of hallucinating visual features to learn ZSL/GZSL classifier.

doi:10.6342/NTU202301924

Algorithm 2: Hallucinating Visual Features for (Generalized) Zero-Shot Recognition

```
Input: Attribute encoder $\mathcal{F}_a$, visual encoder $\mathcal{F}_v$, semantic-conditioned
       transformation $\mathcal{T}$, generator $\mathcal{G}$, and KNN Classifier $\mathcal{C}$
Data: Seen-class data $\mathcal{D}_S = \{(x, a, y) | x \in \mathcal{X}_S, a \in \mathcal{A}, y \in \mathcal{Y}_S\}$ and
      Unseen-class data $\mathcal{D}_U = \{(a, y) | a \in \mathcal{A}, y \in \mathcal{Y}_U\}$
Output: Label prediction $y^*$
Randomly sample a minibatch $\{a_s, (x_{j,1}, x_{j,2})\}$ from $\mathcal{D}_S$ and $a_u$ from $\mathcal{D}_U$;
$\tilde{a}_u = \mathcal{F}_a(a_u)$;
$z_{j,1} = \mathcal{F}_v(x_{j,1}), z_{j,2} = \mathcal{F}_v(x_{j,2})$;
$\Delta z_j = f_{\text{diff}}(z_{j,1}, z_{j,2})$;
$\Delta z_{j \to u} = \mathcal{T}_{j \to u}(\Delta z_j)$;
$z \sim \mathcal{N}(0, I)$;
$\tilde{z}_u = f_{\text{add}}(\Delta z_{j \to u}, z)$;
$\hat{x}_u = \mathcal{G}(\tilde{a}_u, z)$, $\tilde{x}_u = \mathcal{G}(\tilde{a}_u, \tilde{z}_u)$;
if Zero-Shot Recognition then
    $\mathcal{C} \xleftarrow{\text{construct KNN classifier}} \hat{x}_u, \tilde{x}_u$;
    $y_u^* = \arg\max_{y \in \mathcal{Y}} p(y | x_{u,test}; \mathcal{C})$;
else
    Generalized Zero-Shot Recognition
    $\tilde{a}_s = \mathcal{F}_a(a_s)$;
    $\tilde{z}_s = f_{\text{add}}(\Delta z_{j \to s}, z)$;
    $\hat{x}_s = \mathcal{G}(\tilde{a}_s, z)$, $\tilde{x}_s = \mathcal{G}(\tilde{a}_s, \tilde{z}_s)$;
    $\mathcal{C} \xleftarrow{\text{construct KNN classifier}} \hat{x}_u, \tilde{x}_u, \hat{x}_s, \tilde{x}_s$;
    $y_u^* = \arg\max_{y \in \mathcal{Y}} p(y | x_{u,test}; \mathcal{C})$;
    $y_s^* = \arg\max_{y \in \mathcal{Y}} p(y | x_{s,test}; \mathcal{C})$;
end
```

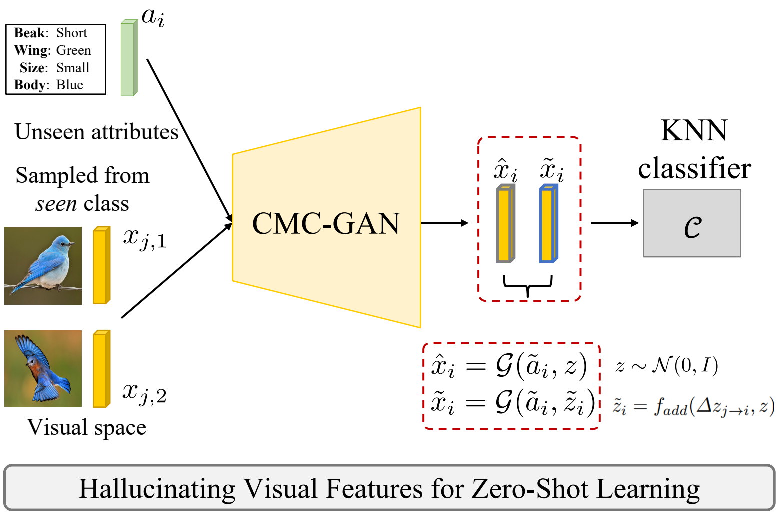

Beak: Short
Wing: Green
Size: Small
Body: Blue
a_i
Unseen attributes
Sampled from *seen* class
$x_{j,1}$
$x_{j,2}$
Visual space
CMC-GAN
$\hat{x}_i$ $\tilde{x}_i$
KNN classifier
$\mathcal{C}$
$\hat{x}_i = \mathcal{G}(\tilde{a}_i, z)$
$\tilde{x}_i = \mathcal{G}(\tilde{a}_i, \tilde{z}_i)$
$z \sim \mathcal{N}(0, I)$
$\tilde{z}_i = f_{add}(\Delta z_{j \to i}, z)$
Hallucinating Visual Features for Zero-Shot Learning

Figure 2.3: Hallucinating Visual features for Zero-Shot Learning. Once our CMC-GAN is learned, visual features of unseen classes are synthesized from the semantic embedding $\tilde{a}$ of unseen classes, together with either randomly sampled noise vector $z \sim \mathcal{N}(0, I)$ or that with intra-class visual diversity transferred from the randomly sampled seen classes (*i.e.*, $\{x_{j,1}, x_{j,2}\}$). With these visual features observed, a KNN classifier is trained accordingly for (generalized) zero-shot learning.

2.4 Experiments

2.4.1 Implementation Details

In all of our experiments, we implement our model using PyTorch and choose ADAM as the optimizer to train our network, with the learning rate, β_1 and β_2 set as 10^{-4}, 0.5, and 0.999, respectively. Our CMC-GAN contains a visual encoder $\mathcal{F}_v$, an attribute encoder $\mathcal{F}_a$, a difference mapping function f_{diff}, an addition mapping function f_{add}, a feature generator $\mathcal{G}$, a discriminator $\mathcal{D}$ with an auxiliary classifier, and an attribute regressor $\mathcal{R}_a$. For $\mathcal{F}_v$, we utilize a single fully-connected (FC) layer to encode the visual features to the visual latent space. The same is also applied to our $\mathcal{F}_a$ for embedding class attributes to semantics features. As for f_{diff} and f_{add}, they are both implemented with a single FC layer to realize the subtraction and addition operations. For the feature generator $\mathcal{G}$, we implement it using multiple fully-connected (FC) layers with 4,096 hidden units. For the discriminator $\mathcal{D}$, we use two FC layers with leaky rectified linear unit (leaky-ReLU) activation functions to realize the two output branches, which produce the binary real/fake prediction and the correct classification result for each visual sample, respectively. Finally, the attribute regressor $\mathcal{R}_a$ consists of two FC layers activated by a leaky-ReLU. We train our network on a single NVIDIA GeForce GTX 1080-Ti GPU with 11 GB memory.

In addition to the details of our network architecture, we further provide the hyperparameters used for each dataset during training. As noted in Section 2.3, the full objective function of our network is

$$\begin{aligned}&\min_{\mathcal{F}_v,\mathcal{F}_a,\mathcal{R}_a,\mathcal{G},f_{diff},f_{add}} \max_{\mathcal{D}} \mathcal{L} \\ &\qquad = \mathcal{L}_{WGAN} + \lambda_1 \mathcal{L}_{CLS} + \lambda_2 \mathcal{L}_{\text{cross}}.\end{aligned} \tag{2.10}$$

In all of our experiments, we set $\lambda_1 = 0.01$ and $\lambda_2 = 1$ for the four benchmark datasets (*i.e.,* CUB, AWA, SUN, FLO), which are determined via cross validation. The sensitivity analysis for λ_1 and λ_2 are also provided in the following.

2.4.2 Datasets and Evaluation Metrics

Datasets

CUB [90] (Caltech-UCSD Birds) is a fine-grained dataset containing a total of 11,788 images from 200 different types of birds, where each bird class corresponds to a semantic vector of 312 attributes (*e.g.,* bill shape, wing color...). It is split into 100 training, 50 validation, and 50 testing classes.

AWA [91] (Animals with Attributes, AWA1) is a coarse-grained dataset containing a total of 30,475 images from 50 animal categories, each corresponding to a semantic vector of 85 attributes (*e.g.,* black, white, stripes, etc.). It is split into 27 training, 13 validation, and 10 test classes.

SUN [92] (Scene Categorization Benchmark) is a fine-grained dataset containing a total of 14,340 images from 717 different types of visual scenes, each associated with a semantic vector of 102 attributes. It is split into 580 training, 65 validation, and 72 testing classes.

FLO [93] (Oxford-Flowers) is a fine-grained dataset containing a total of 8,189 images from 102 different types of flowers (62 for training, 20 for validation, and 20 for testing). Since the attribute information is not available in this dataset, the sentence embedding collected from [94] is applied as the class description.

Evaluation Metrics

Our evaluation protocol calculates per-class top-1 accuracy, where the class with the highest softmax score is selected as the predicted answer. In the ZSL setting, the accuracy of each unseen class is obtained independently, and then averaged over all unseen classes (denoted as U). Apart from U, we also compute the average per-class accuracy of seen classes in the GZSL setting, denoted as S. The final result of GZSL is calculated as the harmonic mean of S and U, that is, $H = (2 \times S \times U)/(S + U)$.

2.4.3 Evaluation and Comparisons

Zero-Shot Learning

We compare our CMC-GAN with state-of-the-art methods for conventional zero-shot learning, and the results are reported in Table 2.2. From this table, we see that our model presented satisfactory top-1 accuracy among state-of-the-art methods. In summary, we achieved 61.4 % on CUB, 71.4 % on AWA, 63.7 % on SUN, and 69.8 % on FLO, respectively. Specifically, TAFE-Net [96], and TCN [97] were recently proposed embedding-based models, boosting the accuracy from the previous works (*e.g.,* [98]) via episodic training strategy or transferable contrastive learning. Other generation-based models [99, 2, 101, 72, 55, 79, 73, 105, 56] synthesized fake samples, and achieved improved results over embedding based methods. While we also adopt GAN as our basic model, we take advantage of our

doi:10.6342/NTU202301924

Table 2.2: Performance comparisons on conventional ZSL in terms of top-1 accuracy (%). The top part is embedding-based methods and the bottom part is generation-based methods.

Method	CUB	AWA1	SUN	FLO
DeViSE [75]	52.0	54.2	56.5	45.9
SJE [52]	53.9	65.6	53.7	53.4
ESZSL [76]	53.9	58.2	54.5	51.0
ALE [67]	54.9	59.9	58.1	48.5
Relation-Net [95]	55.6	68.2	-	-
TAFE-Net [96]	56.9	70.8	60.9	-
TCN [97]	59.5	70.3	61.5	-
LATEM [98]	49.3	55.1	55.3	40.4
GAZSL [99]	55.8	68.2	61.3	60.5
f-CLSWGAN [2]	57.3	68.2	60.8	67.2
C-VAEGAN [100]	54.9	69.9	59.0	-
SE-GZSL [101]	59.6	69.2	63.4	-
Cycle-WGAN [72]	57.8	65.6	59.7	68.6
f-VAEGAN [55]	61.0	71.1	**64.7**	67.7
LisGAN [79]	58.8	70.6	61.7	69.6
F2F [73]	58.5	69.3	61.5	-
Our CMC-GAN	**61.4**	**71.4**	63.7	**69.8**

doi:10.6342/NTU202301924

Table 2.3: Performance comparisons on GZSL in terms of top-1 accuracy (%). Note that U and S denote the accuracy of unseen and seen classes, respectively. The harmonic mean H is calculated by $H = (2 \times S \times U)/(S + U)$. The top part is embedding-based methods and the bottom part is generation-based methods.

Dataset	CUB			AWA1			SUN			FLO		
Method	U	S	H	U	S	H	U	S	H	U	S	H
DeViSE [75]	23.8	53.0	32.8	13.4	68.7	22.4	16.9	27.4	20.9	9.9	44.2	16.2
SJE [52]	23.5	59.2	33.6	11.3	74.6	19.6	14.7	30.5	19.8	13.9	47.6	21.5
ESZSL [76]	2.4	70.1	4.6	5.9	77.8	11.0	11.0	27.9	15.8	11.4	56.8	19.0
CMT [77]	7.2	49.8	12.6	0.9	87.6	1.8	8.1	21.8	11.8	-	-	-
SAE [68]	7.8	54.0	13.6	1.8	77.1	3.5	8.8	18.0	11.8	-	-	-
ALE [67]	4.6	73.7	8.7	14.0	81.8	23.9	21.8	33.1	26.3	13.3	61.6	21.9
LATEM [98]	15.2	57.3	24.0	7.3	71.7	13.3	14.7	28.8	19.5	6.6	47.6	11.5
CADA-VAE [54]	51.6	53.5	52.4	57.3	72.8	64.1	47.2	35.7	40.6	-	-	-
LFGAA [102]	36.2	80.9	50.0	27.0	93.4	41.9	18.5	40.0	25.3	-	-	-
Relation-Net [95]	31.4	91.3	46.7	38.1	61.1	47.0	-	-	-	-	-	-
TAFE-Net [96]	41.0	61.4	49.2	50.4	84.4	63.2	27.9	40.2	33.0	-	-	-
TCN [97]	52.6	52.0	52.3	49.4	76.5	60.0	31.2	37.3	34.0	-	-	-
SYNC [103]	7.4	66.3	13.3	10.0	90.5	18.0	7.9	43.3	13.4	-	-	-
GAZSL [99]	23.9	60.6	34.3	19.2	86.5	31.4	21.7	34.5	26.7	28.1	77.4	41.2
f-CLSWGAN [2]	43.7	57.7	49.7	57.9	61.4	59.6	42.6	36.6	39.4	59.0	73.8	65.6
C-VAEGAN [100]	42.7	45.6	44.1	62.7	60.6	61.6	44.4	30.9	36.5	-	-	-
SE-GZSL [101]	41.5	53.3	46.7	58.3	68.1	62.8	40.9	30.5	34.9	-	-	-
Cycle-WGAN [72]	46.0	60.3	52.2	56.4	63.5	59.7	48.3	33.1	39.2	59.1	71.1	64.5
f-VAEGAN [55]	48.4	60.1	53.6	57.6	70.6	63.5	45.1	38.0	41.3	56.8	74.9	64.6
LisGAN [79]	46.5	57.9	51.6	52.6	76.3	62.3	42.9	37.8	40.2	57.7	83.8	68.3
F2F [73]	47.0	54.8	50.6	57.3	67.1	61.8	45.3	36.8	40.6	-	-	-
DASCN [80]	45.9	59.0	51.6	59.3	68.0	63.4	42.4	38.5	40.3	-	-	-
SGAL [104]	40.9	55.3	47.0	52.7	74.0	61.5	35.5	34.4	34.9	-	-	-
OCD-CVAE [105]	44.8	59.9	51.3	59.5	73.4	65.7	44.8	42.9	43.8	-	-	-
LsrGAN [56]	48.1	59.1	53.0	54.6	74.6	63.0	44.8	37.7	40.9	-	-	-
Our CMC-GAN	52.6	65.1	**58.2**	63.2	70.6	**66.7**	48.2	40.8	**44.2**	64.5	80.2	**71.5**

doi:10.6342/NTU202301924

Table 2.4: Ablation studies on the design of the proposed CMC-GAN on three benchmark datasets. Note that z denotes the input randomly sampled from $\mathcal{N}(0, I)$, Δz_j is the extracted visual diversity from class j, and $\Delta z_{j\to i}$ represents that translated from class j to i via semantic-conditioned transformer $\mathcal{T}$. The bold numbers indicate the best results.

	Objectives and components						CUB			AWA1			SUN		
	$\mathcal{L}_{CLS}$	z	Δz_j	$\Delta z_{j\to i}$	$\mathcal{L}^A_{con}$	$\mathcal{L}^V_{con}$	U	S	H	U	S	H	U	S	H
Baseline [2]	✓	✓	×	×	×	×	43.7	57.7	49.7	57.9	61.4	59.6	42.6	36.6	39.4
Ours w/o $\mathcal{T}, \mathcal{L}^V_{con}, \mathcal{L}^A_{con}$	✓	✓	✓	×	×	×	44.5	54.5	49.0	56.8	67.1	61.5	45.0	35.9	39.9
Ours w/o $\mathcal{L}^V_{con}, \mathcal{L}^A_{con}$	✓	✓	✓	✓	×	×	52.2	55.0	53.5	61.4	66.1	63.7	45.6	38.4	41.7
Ours w/o $\mathcal{L}^V_{con}$	✓	✓	✓	✓	✓	×	51.0	58.0	54.3	62.1	66.7	64.3	45.8	38.7	41.9
Ours w/o $\mathcal{L}^A_{con}$	✓	✓	✓	✓	×	✓	52.4	58.1	55.1	60.4	67.6	63.8	46.2	39.1	42.4
Ours (CMC-GAN)	✓	✓	✓	✓	✓	✓	52.6	65.1	**58.2**	63.2	70.6	**66.7**	48.2	40.8	**44.2**

designed semantics-guided intra-category knowledge transfer for augmenting class-specific data. Note that, the SUN dataset contains complex outdoor visual scenes with relatively limited attributes (*i.e.,* 102 dimensions), causing the intra-class variation to be more difficult to model. Nevertheless, our network still achieved the second-best result on SUN, and performed best on three out of four datasets. Thus, the effectiveness of our model for conventional zero-shot recognition can be successfully verified.

Generalized Zero-Shot Learning

We now provide comparisons with the state-of-the-art approaches in generalized zero-shot learning as reported in Table 2.3. The setting of GZSL is more challenging than conventional ZSL due to a model trained only on seen data would easily lead the unseen objects to be classified into seen classes, resulting in biased problem. For the purpose of measuring the performance on both seen and unseen classes, we follow [64] and adopt harmonic mean, which prevents the impact of extreme values, to evaluate the performance of GZSL. As shown in Table 2.3, we performed favorably against state-of-the-art methods, achieving promising performances in terms of harmonic mean (*i.e.,* 58.2 % on CUB, 66.7 % on AWA,

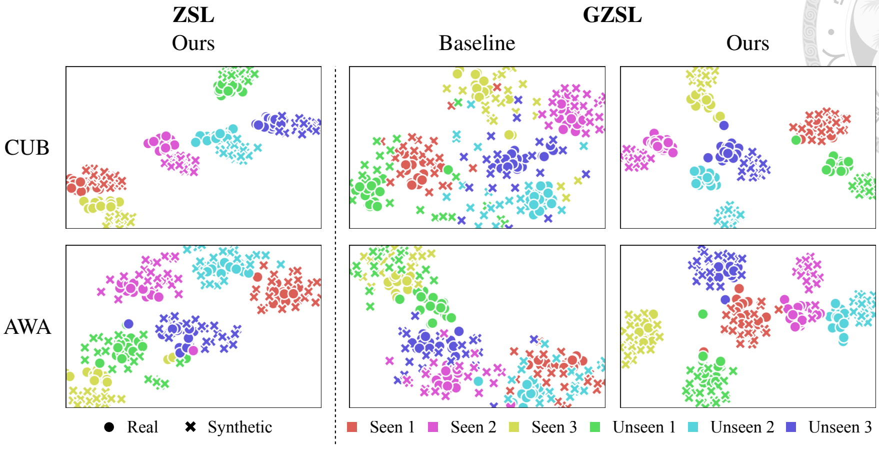


Figure 2.4: t-SNE visualization of hallucinated data for CUB and AWA. For ZSL, samples of 6 unseen classes are selected (150 samples for each class). For GZSL, we sample 3 seen classes and 3 unseen classes for visualization, and compare the results produced by the baseline approach [2]. For both ZSL and GZSL, we see that our model produces data with improved diversity while not overfitting those of seen classes. Note that real data of unseen classes are additionally shown in this figure for visualization and comparison purposes.

44.2 % on SUN, and 71.5 % on FLO). This confirms that our model is capable of exploiting class-specific diversity for synthesizing the pseudo training samples. In particular, as shown in the top rows in Table 2.3, embedding-based works generally suffered from the aforementioned biased problems and reported poor results. Thus, their model cannot be generalized to unseen classes. Though other recent approaches [54, 96, 97] provided favorable results in conventional ZSL, they presented degraded performances in GZSL due to the inherent heterogeneous gap across distinct modalities, that hampers the ability of generalization across seen and unseen classes. As shown in the bottom rows in Table 2.3, we observe that generation-based methods [2, 72, 100, 55, 79, 80, 105, 56] reported significant performance drops under GZSL settings. For instance, Cycle-WGAN [72] and DASCN [80] follow f-CLSWGAN [2] and apply semantic consistency to enforce the generated visual features with proper semantics that generates outputs simply

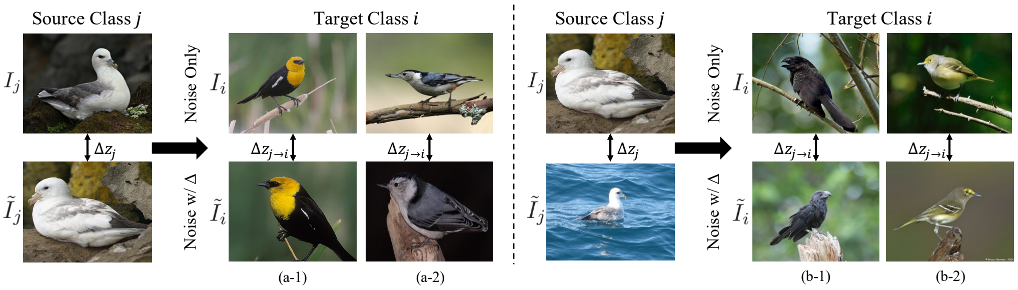


Figure 2.5: Example class-specific diversity transfer from a seen class j to unseen class i (*i.e.,* Δz_j to $\Delta z_{j\to i}$).

from class attributes with random noise. In addition, the recent state-of-the-art LsrGAN [56] leverages the semantic relationships between seen and unseen classes to guide the synthesized visual features retaining the relationships observed from semantic space. However, such methods do not explicitly model intra-category knowledge for visual feature hallucination, and thus cannot ensure the synthesized visual features are sufficiently diverse.

2.4.4 Analysis of Cross-Modal Consistency GAN

Ablation Study

We now conduct the ablation study in Table 2.4 to verify our network design (more parameter analysis is available in supplementary material). We consider f-CLSWGAN [2] as the baseline model to start with, which generates fake visual features simply conditioned on the class attribute with random noise. In this table, we consider the performances of our CMC-GAN 1) without transformation layer $\mathcal{T}$ and cross-modal consistency (*i.e.,* $\mathcal{L}^V_{con}$ and $\mathcal{L}^A_{con}$), 2) without $\mathcal{L}^V_{con}$ and $\mathcal{L}^A_{con}$, 3) without semantics consistency $\mathcal{L}^A_{con}$, and 4) without visual diversity consistency $\mathcal{L}^V_{con}$. For each model, we evaluate the performance of GZSL in harmonic mean. From the results shown in this table, our model surpassed other controlled versions and the baseline on all three benchmark datasets. We note that, without the presence of the transformation layer and cross-modal consistency, the harmonic

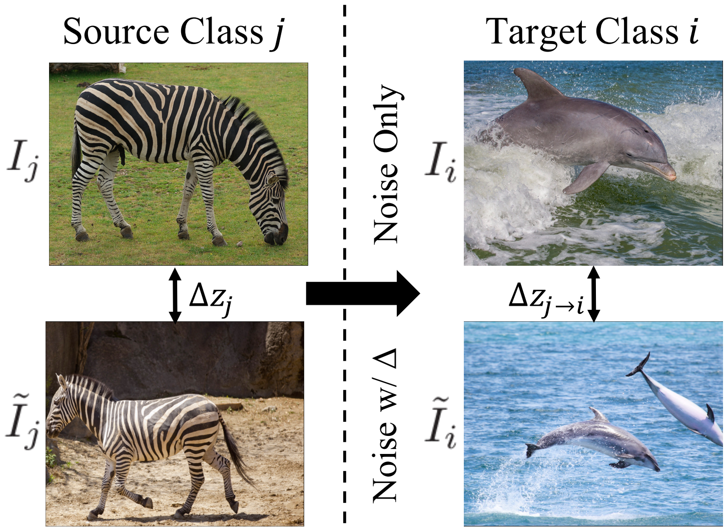


Figure 2.6: Example failure case of class-specific diversity transfer from a seen class j to unseen class i.

mean of top-1 accuracy may drop significantly even below or near the baseline. This is because that, without $\mathcal{T}$ translating the class-specific information to the target class, the derived visual difference Δz_j only fits the source data distribution, resulting in overfitting on the seen classes. With $\mathcal{T}$ deployed, the improvement on three datasets can be observed (comparison with the second and third row). The above experiments confirm the effectiveness of our semantics-guided intra-category knowledge transformer.

Moreover, we see that the performance would drop when the cross-modal consistency is disabled. This is due to the fact that, without observing such consistency, we cannot encourage the preservation of semantics and visual diversity. In fact, disregarding either semantics or visual diversity consistency would not be desirable for our model, since the semantics mismatch or mode collapse would occur. With the above experiments provided, we can verify the effectiveness of each module deployed in our CMC-GAN.

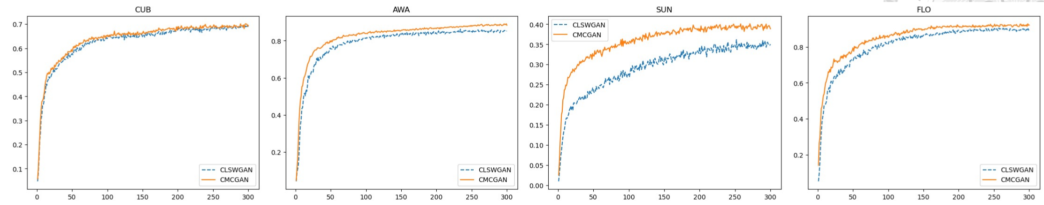


Figure 2.7: Convergence of the top-1 accuracy in terms of the number of epochs for the generated training samples from the seen classes for CUB, AWA, SUN, and FLO.

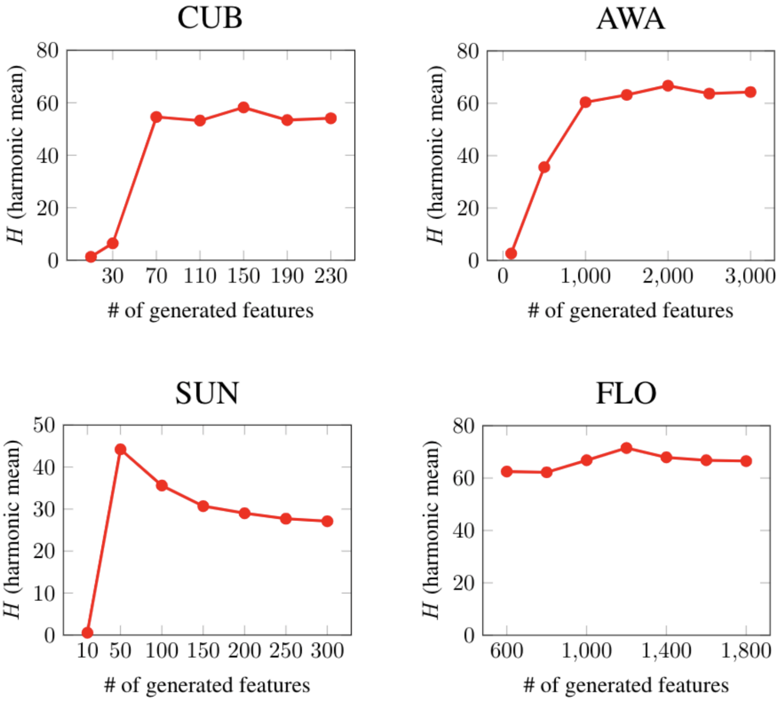


Figure 2.8: The impact of different numbers of synthetic visual samples per category. Note that x and y axes indicate the number of generated features and harmonic mean (H) respectively.

Visualization

We now qualitatively assess the ability of our CMC-GAN to represent semantic and visual information for the synthesized data. As shown in Fig. 2.4, we visualize the visual features $\tilde{x}$ synthesized on CUB and AWA datasets using t-SNE. The left part of Fig. 2.4 illustrates the generated visual data of five unseen categories produced by our model, which were properly separated and sufficiently diverse. When comparing to the baseline model [2], we consider the GZSL setting and present the results in the right-hand side of Fig. 2.4. It can be seen that, the synthesized data of [2] suffer from severely biased problems (*i.e.,* data of unseen categories are visually similar to those of seen classes). On the other hand, visual data of unseen classes produced by CMC-GAN still exhibited satisfactory inter-class separation with sufficient intra-class diversity.

To further demonstrate our ability to translate class-specific diversity across different categories, Fig. 2.5 shows the exampled unseen-class results synthesized from seen classes. Specifically, we randomly select an image pair $(I_j, \tilde{I}_j)$ from a seen class j, and transform its intra-class diversity Δz_j to another unseen class i. Since our proposed model only hallucinates visual *features*, the *images* (*i.e.,* $(I_i, \tilde{I}_i)$) shown in Fig. 2.5 are selected by those whose image features are *closest* to the hallucinated ones (from the same category i). From the results shown in this figure, we observe that our model is able to model and convert intra-class diversity into the unseen class in terms of visual concepts such as pose (right→left) and size (small→large) variations, as shown in columns (a-1), (b-1) and (a-2), (b-2) in Fig. 2.5, respectively. This confirms the effectiveness of our semantics-guided intra-category knowledge transfer module, and our CMC-GAN in producing unseen-class outputs with desirable semantic and visual information.

In Fig. 2.6, we show failure diversity transfer cases by our CMC-GAN. As evident in this figure, to transfer the intra-class variation (*e.g.,* pose) from "zebra" to "dolphin", whose appearance bears little resemblance to "zebra", is more challenging than the transfer across fine-grained classes (*e.g.,* different birds in CUB dataset). One possible solution to boost the transfer ability across more coarse-grained datasets is to apply ad-hoc techniques like [83] or meta-class information for selecting proper classes and their variants to transfer. We leave this among future research directions.

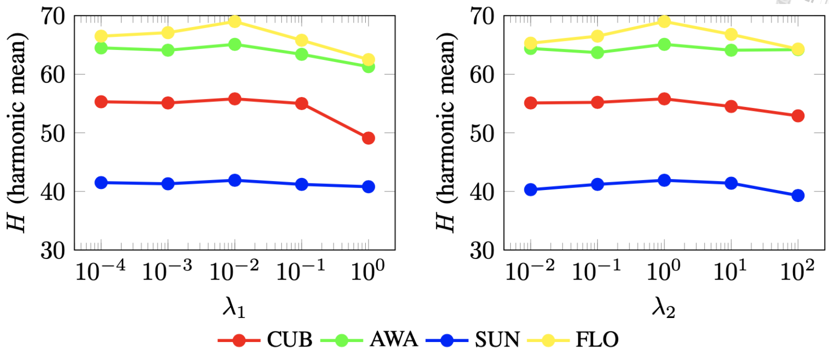


Figure 2.9: Sensitivity analysis for hyperparameters λ_1 and λ_2. Note that x and y axes indicate the value of λ and harmonic mean (H) respectively.

2.4.5 Parameter Analysis

In this section, we first evaluate the impact of the number of generated visual features, and then provide the sensitivity analysis for the hyperparameters.

Efficiency of the visual feature hallucination

We measure the seen class accuracy of the classifier trained on generated features of seen classes w.r.t. the training epochs and compare with the baseline method f-CLSWGAN [2] in Fig. 2.7 to evaluate the efficiency of visual feature hallucination process and also the quality of synthesized visual features. As shown in Fig. 2.7, our CMC-GAN converges fast and achieves higher accuracy against f-CLSWGAN [2].

Impact of the number of generated visual features

We evaluate how the number of synthetic visual samples per class impacts the performance in the GZSL setting. As illustrated in Fig. 2.8, we observe that the harmonic mean (H) is improved when the number of synthetic features is increased, but saturates in a certain degree in the results of CUB, AWA, and FLO. We note that the performance in SUN grows first and then decays with the increasing number of synthetic samples. The reason is that SUN dataset contains only around 20 images

doi:10.6342/NTU202301924

per class with relatively limited attributes (*i.e.,* 102 dimensions) for describing complex visual scenes. Hence, too many generated samples would cause severely noisy pseudo samples, leading to degraded performance. We determined the final number of synthetic visual samples for each dataset according to the experiment performed in this section. (*i.e.,* 150 on CUB, 2000 on AWA, 50 on SUN, and 1200 on FLO).

Sensitivity analysis for hyperparameters

The hyperparameters of our model are tuned via cross validation. In this section, we conduct a detailed analysis of the sensitivity of hyperparameters λ_1 and λ_2. As shown in Fig. 2.9, the performance does not exhibit drastic fluctuations despite using different sets of λ_1 and λ_2, showing further that our model is stable and robust and those hyperparameters are not the most influential factors to the result.

2.5 Conclusion

In this thesis, we proposed a data hallucination-based model of Cross-Modal Consistency GAN (CMC-GAN) for ZSL/GZSL. In order to hallucinate visual data of unseen categories, our model performs semantics-guided intra-category knowledge transfer, which translates visual diversity across image categories under the guidance of the associated semantic features. To alleviate possible mode collapse and biased problems during hallucination, our model observes attribute and visual consistencies at the associated levels, ensuring that the synthesized data would sufficiently represent the category of interest. Finally, we conducted experiments on four benchmark datasets, which quantitatively and qualitatively support the effectiveness of our CMC-GAN over the state-of-the-art embedding and hallucination-based methods for ZSL and GZSL.

doi:10.6342/NTU202301924

Chapter 3

Knowledge Transfer for Unseen Domains

Domain generalization (DG) aims to transfer the learning task from a single or multiple source domains to unseen target domains. To extract and leverage the information which exhibits sufficient generalization ability, we propose a simple yet effective approach of Adversarial Teacher-Student Representation Learning, with the goal of deriving the domain generalizable representations via generating and exploring out-of-source data distributions. Our proposed framework advances Teacher-Student learning in an adversarial learning manner, which alternates between knowledge-distillation based representation learning and novel-domain data augmentation. The former progressively updates the teacher network for deriving domain-generalizable representations, while the latter synthesizes data out-of-source yet plausible distributions. Extensive image classification experiments on benchmark datasets in multiple and single source DG settings confirm that, our model exhibits sufficient generalization ability and performs favorably against state-of-the-art DG methods.

doi:10.6342/NTU202301924

3.1 Introduction

Deep neural networks have achieved promising performance on a wide variety of tasks. However, these networks assume the training and testing samples fall in the same data distribution. Such a strong assumption would limit the applicability of the learned models in real-world scenarios (e.g., cross-city autonomous driving or multi-cite medical imaging task), in which training and testing data are typically observed under different conditions. In other words, the generalizability of the model at *unseen* target domains might be poor due to *unexpected domain shifts*. To tackle the domain discrepancy, domain generalization (DG) has been proposed and drawn increasing attention recently.

The aim of DG is to train models using data observed from single or multiple source domains, while expecting that the model would be generalized to unseen target domains. Most existing DG approaches focus on deriving domain-invariant features [106] among multiple source domains or adopting meta-learning techniques [107, 108, 109, 110], which would simulate domain shifts during the meta-training stage. However, the features derived by the above methods are generally guaranteed to be invariant to the seen source domains, not the generalizability of the learned representation to describe unseen domain data. To overcome the limitation, [111, 3, 112, 113] turn to leverage data generation techniques for diversifying the source distributions, and thus avoid overfitting on source domains yet improve the generalization ability of models. Specifically, several works [111, 3, 112] choose to generate novel-domain images by either perturbing the style of source data to confuse the domain classifier [111, 3], or transporting the source data to novel styles via optimal transport based objective [112]. [113] adopts Mixup [114] to interpolate the feature statistics between samples from different domains. While the above methods perform well, designing an objective for generating samples with DG guarantees remains a challenging and open problem.

Recently, self-supervised pre-training manifests the potential to derive generalizable representation, which serves as a promising start point for downstream

doi:10.6342/NTU202301924

tasks (e.g., image segmentation or object detection). In domain generalization, a number of self-supervision techniques have been introduced [115, 116, 117] to improve network transferability by discovering the intrinsic properties within images. For instance, [115, 116, 117] adopt jigsaw puzzles as the pretext task, which predicts the relative positions of image patches to constrain the semantic feature learning in a multi-task training fashion. Recently, contrastive learning approaches [118, 119, 120, 121, 122] have been proposed and widely applied, which establish the representation learning from multiple views of an image to extract the task-relevant information and discard task-irrelevant noise. However, the concept of such multiview learning [123, 122] is simply realized by hand-crafted image transformations (e.g., *RandomResizedCrop, Color Jittering, or Gaussian Blur*). The effectiveness of these hand-crafted image transformations for benefiting the generalization to unseen distributions is still not guaranteed.

In this thesis, we propose a unique *Adversarial Teacher-Student Representation Learning* framework for tackling domain generalized visual classification. Based on the recent success of contrastive learning, we advance the concept of multi-view learning into DG regime for augmenting source instances to out-of-source styles and diversifying training distributions. To be more precise, with the goal of learning representations which are robust to unseen domain shift, we propose to jointly perform *Domain Generalized Representation Learning* and *Novel Domain Augmentation* in an adversarial learning manner. Based on Teacher-Student learning schemes [124, 120, 125], our framework utilizes original images as inputs to the teacher network and takes stylized augmentations as input to the student network. To ensure both learning stages produce domain generalized representation, we adopt the Teacher-Student co-training scheme, which progressively refines Teacher by the distilled knowledge learned from Student by observing augmented novel-domain data, enabling Teacher to be generalizable to data with out-of-source distributions. On the other hand, *Adversarial Novel Domain Augmentation* aims at augmenting unseen domain data using source-domain training

doi:10.6342/NTU202301924

instances. The objective is to *maximize* the discrepancy between the input and augmented data, derived from the teacher and student modules, respectively. In order to have such augmented data exhibit sufficient domain differences, the above discrepancy will be calculated using features derived from data across different source domains. By iteratively training the above two stages in an adversarial learning fashion, the resulting model (Teacher) would be able to derive domain generalizable representations.

The contributions of this thesis are highlighted as below:

- Different from existing meta-learning based approaches, we propose a teacher-student adversarial learning scheme for addressing domain generalization classification problems.

- In the stage of Domain Generalized Representation Learning, the student network observes augmented novel-domain data and distills the information to update the teacher network, allowing derivation of domain generalizable representation.

- In the stage of Novel Domain Augmentation, the generator aims at producing unseen yet plausible domain data, which maximizes the discrepancy between augmented and existing domains while the semantic information is preserved.

- Evaluations on several benchmark datasets in multiple and single source domain settings verify that our method performs favorably against existing DG approaches and exhibits sufficient domain generalization capability.

3.2 Related Works

Domain Generalization (DG). Different from domain adaptation (DA), which observes both source and target-domain training data for performing learning tasks across domains [14, 16, 11, 126, 127], DG deals with a more realistic yet challenging setting. More precisely, DG aims at generalizing the model trained only on single or multiple source domains to recognize the test instance in unseen but similar target domain. With only source-domain data observed during training, a number of works [108, 107, 128] apply meta-learning for learning domain-invariant features. These methods typically partition source domains into meta-train and meta-test splits to simulate the domain shifts during training. Feature-Critic [128] meta-learns a critic network to evaluate the generalized degree of extracted features for encouraging robust feature derivation. [109] introduces an episodic training that cross-trains domain-specific feature extractors and classifiers to let the learned model invariant to the domain shift. MLDG [107] and MASF [110] both adopt gradient based meta-learning to simulate the domain shift, while [110] additionally enforces local and global constraints in meta-training. In addition to meta-learning approaches, [115] jointly solves jigsaw puzzle as an auxiliary task with standard classification in a multi-task fashion. RSC [129] iteratively discards the dominant features on the training data to improve generalization. Nevertheless, these approaches employ solely limited source domains to derive generalizable features, which still draws a concern about over-fitting on source domains [112, 130] and restricts the generalization ability to unseen domains.

Recent research works consider data generation as an alternative technique for domain generalization, which increase the diversity of training data distribution. To achieve this goal, [111, 3] are inspired by adversarial attack [131]. CrossGrad [111] perturbs source data by adding adversarial gradients; DDAIG [3] learns a transformation network that outputs perturbations to confuse the domain classifier. However, such perturbed images do not necessarily exhibit sufficient data domain diversity. In contrast of adding perturbation to images, L2A-OT [112] learns a conditional generator that transforms images from a source distribution to a pseudo-novel distribution by an optimal transport based objective. MixStyle [113] produces image features with mixed feature statistics across source domains. Very recently, PDEN [132] utilizes a progressive learning strategy for single-source domain generalization, which iteratively expands the training data set by adding

doi:10.6342/NTU202301924

augmented data. Note that although they adopt contrastive and adversarial learning objectives which are similar to ours, our proposed approach is able to tackle both multi-source and single-source DG problems, and also comes with superior memory efficient performance and comparably stable training process.

Self-Supervised Learning (SSL). Self-supervision is a recent paradigm for unsupervised learning. The idea is to design pretext tasks for feature learning to facilitate the downstream task learning. Such auxiliary pretext tasks can be predictions of the image colors [133], relative locations of patches from the same image [115, 116], and image rotation [134]. Very recently, contrastive learning [118, 119, 120, 121, 122] has achieved promising results on network pretraining to learn generalized image features. [123, 122] reveal that the success of contrastive learning is typically built on the multi-view perspective, and prove theoretically and experimentally that the compact and robust representations can be learned by deriving the invariance among multiple views of an image. We adapt the above concept of multi-view learning into DG regime. We focus on learning novel-domain data augmentations across source-domain instances in an adversarial training fashion. As detailed and verified later, our proposed learning scheme would produce domain generalizable representation for unseen target-domain data, and performs favorably against state-of-the-art DG approaches.

3.3 Proposed Method

3.3.1 Problem Formulation and Model Overview

For the sake of completeness, we first define the problem setting and notations used in this thesis. We assume that training data are observed from N source domains $\mathcal{D}_{tr} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$, each of which contains a set of image and label pairs $\mathcal{D}_i = \{X_i, Y_i\}$. Our goal is to learn a model which would exhibit sufficient generalization capability, so that classification of test data in unseen

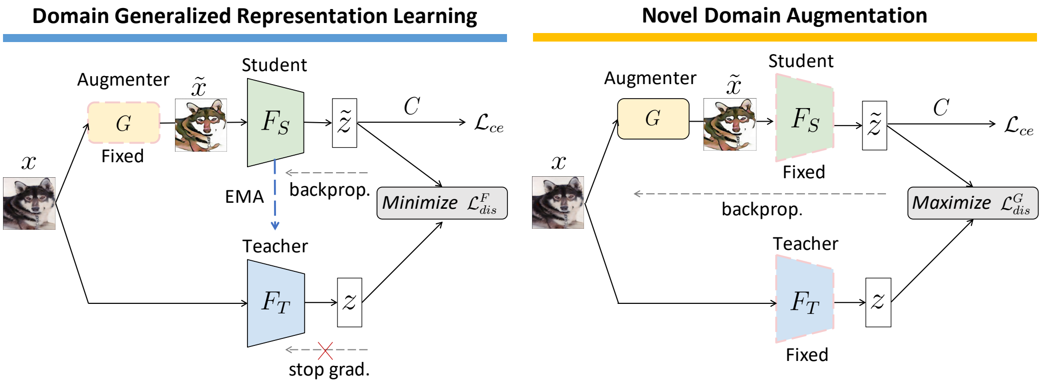


Figure 3.1: Overview of our Adversarial Teacher-Student Representation Learning scheme, which includes the teacher network F_T, the student network F_S, classifier C, and novel-domain augmenter G. Note that we alternate between the stages of domain generalized representation learning and novel-domain augmentation in a mutually beneficial manner, resulting in discriminative yet domain generalized representations.

target domains can be performed. In order to derive domain-generalized feature representations, we present a novel *Adversarial Teacher-Student Representation Learning* framework, which is a min-max deep learning framework alternating between the following two stages: *domain generalized representation learning* (Sec. 3.3.2) and *novel domain augmentation* (Sec. 3.3.3), as depicted in Fig. 3.1. For *domain generalized representation learning*, we learn a domain-generalized teacher network (Teacher) F_T with the help from a student network (Student) F_S, which observes synthesized *novel*-domain augmentation and distills knowledge to Teacher. As for *novel-domain augmentation*, the novel-domain augmenter G is learned to observe the discrepancy of Teacher-Student encoders, which progressively generates *strong* novel stylized augmentations to diversify training distributions. Once the learning of the above framework is complete, the teacher network would extract domain-generalized features for the task network (e.g., classifier), and thus classification of unseen target-domain data can be performed accordingly. We now detail our proposed learning schemes in the following subsections.

3.3.2 Teacher-Student Domain Generalized Representation Learning

While techniques based on learning across multiple source domains for DG exist (e.g., using meta-learning techniques like [108, 109, 110]), it is not clear how the learned model and feature representations would be generalized to unseen target domains. Instead of directly fitting models across source domains, we propose *Domain Generalized Representation Learning* based on the Teacher-Student learning scheme, with the goal of extracting domain generalizable feature representations. To ensure our teacher encoder to gain generalizability by observing out-of-source domain information, we deploy a Student F_S for exploring novel-domain augmentation synthesized from the novel-domain augmenter G, while distilling the associated knowledge to update F_T.

To address this representation learning task, we first train the teacher module together with a single-layer classifier C using multiple source-domain data. The standard cross-entropy loss $\mathcal{L}_{ce}$ is utilized to initialize F_T as warm-up. As illustrated in Fig. 3.1, we then input training images x sampled from the source domains into the novel-domain augmenter G (detailed in the following sub-section), producing the style (or domain) perturbed augmentation $\tilde{x}$ yet preserving its semantic information. While such a domain augmented $\tilde{x}$ would be fed into the student module resulting in feature $\tilde{z} = F_S(\tilde{x})$, we also feed the original input x into Teacher to derive $z = F_T(x)$. To ensure that $\tilde{z}$ would contain the same semantic information as z does, we particularly propose an objective to **minimize** the discrepancy between $\tilde{z}$ and z. To be more specific, we define the discrepancy loss $\mathcal{L}^F_{dis}$ to minimize the distance between the normalized features $\tilde{z}$ and z:

$$\min_{F_S} \mathcal{L}^F_{dis}(z, \tilde{z}) = \left\| \frac{z}{\|z\|_2} - \frac{\tilde{z}}{\|\tilde{z}\|_2} \right\|_2^2 = \left\| \frac{F_T(x)}{\|F_T(x))\|_2} - \frac{F_S(\tilde{x})}{\|F_S(\tilde{x})\|_2} \right\|_2^2. \tag{3.1}$$

In addition, we calculate the cross-entropy loss on the domain-augmented feature $\tilde{z}$, i.e., $\mathcal{L}_{ce}(C(\tilde{z}), y)$, which further enforces the classification capability of the student module (note that C indicates the single-layer classifier, and y denotes

doi:10.6342/NTU202301924

the corresponding class label). We note that, in this representation learning stage, only the student network F_S is updated by the above two objectives $\mathcal{L}^F_{dis}(z, \tilde{z})$ and $\mathcal{L}_{ce}(C(\tilde{z}, y)$, and we apply a stop-gradient strategy to forbid F_T and G from being updated by gradients. Thus, optimization of F_S with learning rate γ can be expressed as follows:

$$\theta_S \leftarrow \theta_S - \gamma \frac{\partial(\mathcal{L}^F_{dis}(z, \tilde{z}) + \mathcal{L}_{ce}(C(\tilde{z}), y))}{\partial \theta_S}. \tag{3.2}$$

As for the teacher network F_T, we adopt exponential moving average (EMA) [124, 120, 125] to progressively refine the associated model parameter θ_T. That is, the learned knowledge from Student's parameter θ_S is distilled to update θ_T as follows,

$$\theta_T \leftarrow \tau\theta_T + (1 - \tau)\theta_S, \quad \text{where}\, \tau \in [0, 1), \tag{3.3}$$

Note that τ controls the updates on the teacher network. Finally, it is also worth pointing out that, such a refinement strategy would avoid the teacher module from directly observing unrealistic domain augmentations, which might degrade its domain generalization capability.

3.3.3 Adversarial Novel Domain Augmentation

To motivate the student network to explore sufficient diversity of domain augmentation, we present an adversarial learning scheme, which would progressively perform novel-domain data augmentation in our proposed framework. Inspired by both adversarial learning strategy [23] and multiview learning from SSL [123, 122], we formulate our novel-domain augmentation stage together with representation learning (Sec. 3.3.2) into an adversarial learning framework. As depicted in the right-hand side of Fig. 3.1, we aim at training the novel-domain augmenter G and freezing both F_T and F_S, while the discrepancy between z and $\tilde{z}$ serves as the adversarial guidance. That is, when the above discrepancy is small (i.e., the outputs of Teacher and Student are similar), it implies that such domain augmentations have been seen by existing source-domain data. To encourage more the augmented data to be sufficiently distinct in terms of domain information, we train our novel-domain augmenter by **maximizing** the discrepancy as follows,

$$\max_G \mathcal{L}_{dis}^{G}(z,\tilde{z}) = [\left\| \frac{z}{\|z\|_2} - \frac{\tilde{z}}{\|\tilde{z}\|_2} \right\|_2^2 - m]_- = [\left\| \frac{F_T(x)}{\|F_T(x))\|_2} - \frac{F_S(G(x))}{\|F_S(G(x))\|_2} \right\|_2^2 - m]_-, \tag{3.4}$$

where $[\cdot]_- = min(\cdot, 0)$, and the margin m can either be calculated by the means/centroids of data from each source domain in a mini-match, followed by averaging the L2 distances between the above centroid pairs, or simply viewed as a hyperparameter. It is worth pointing out that, this margin serves as a regularization observed from the separation between existing source domains. Thus, it reflects the desirable domain gap between the augmented and existing domain data.

To guarantee the produced domain augmentations preserve the original categorical content, we still observe the cross-entropy loss $\mathcal{L}_{ce}(C(\tilde{z}), y)$ with regard to $C(\tilde{z})$ and the corresponding label y. Thus, optimization of G can be performed as follows,

$$\theta_g \leftarrow \theta_g - \gamma \frac{\partial(-\mathcal{L}_{dis}^{G}(z,\tilde{z}) + \mathcal{L}_{ce}(C(\tilde{z}), y))}{\partial \theta_g}. \tag{3.5}$$

Note that, we only pretrain the classifier using source domain data available, and keep it fixed during the learning of our teacher-student augmentation framework. If we allow the update of this classifier during the training process, it might observe undesirable outputs and affect the learning of both augmenter and teacher/student modules in the early training stage, where either the augmented data or its extracted features are not yet qualified.

Once the learning of the proposed framework is complete (i.e., alternative optimization between the two stages), we deploy the derived domain generalized Teacher to extract discriminative and transferable features, so that classification of unseen target domain can be performed accordingly. The pseudo code of our Adversarial Teacher-Student Representation Learning is summarized in Algorithm 3.

doi:10.6342/NTU202301924

Algorithm 3: Adversarial Teacher-Student Representation Learning

```
Input: Number of iterations $N_{iter}$, number of warm up iterations $N_{warm}$,
       learning rate $\gamma$, Teacher $F_T$, Student $F_S$, novel-domain augmenter
       $G$ and classifier $C$
Data: $N$ source domains $\mathcal{D}_{tr} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$
Output: Teacher $F_T$
for $i$ in $1 : N_{iter}$ do
    Randomly sample a minibatch $(x, y)$ from source domains ;
    if $i$ $< N_{warm}$ then
        Update $F_T$ and $C$ with $\mathcal{L}_{ce}(C(F_T(x)), y)$;
    else
        Domain Generalized Representation Learning
        $\tilde{x} = G(x)$;
        $z = F_T(x)$, $\tilde{z} = F_S(\tilde{x})$;
        Compute $\mathcal{L}^F_{dis}$ (Eq.1) and $\mathcal{L}_{ce}(C(\tilde{z}), y)$;
        Update $F_S$ via back propagation.
         $\theta_S \leftarrow \theta_S - \gamma \frac{\partial(\mathcal{L}^F_{dis}(z,\tilde{z}) + \mathcal{L}_{ce}(C(\tilde{z}),y))}{\partial\theta_S}$ (Eq.2);
        Update $F_T$ via EMA. $\theta_T \leftarrow \tau\theta_T + (1-\tau)\theta_S,$   where $\tau \in [0, 1)$
         (Eq.3);
        Novel Domain Augmentation
        Compute $\mathcal{L}^G_{dis}$ (Eq.4) and $\mathcal{L}_{ce}(C(\tilde{z}), y)$;
        Update $G$ via back propagation.
         $\theta_g \leftarrow \theta_g - \gamma \frac{\partial(-\mathcal{L}^G_{dis}(z,\tilde{z}) + \mathcal{L}_{ce}(C(\tilde{z}),y))}{\partial\theta_g}$ (Eq.5);
    end
end
```

3.4 Experiments

3.4.1 Datasets and Evaluation Protocol

Datasets. We evaluate our method on several public benchmark datasets. **PACS** [135] is composed of four data domains, *Photo*, *Art painting*, *Cartoon* and *Sketch*, with diverse image colors and styles. Each domain contains 7 categories, with 9991 images in total. Following the experimental protocol proposed by [135], images from source domains are divided into the training split and the validation split, at a ratio of about 9:1. **Office-Home** [136] is comprised of four domains, *Art*, *Clipart*, *Product* and *Real world*, and exists larger label sets of 65 categories, with about 15500 images in total. The dataset contains images of everyday objects with various styles, backgrounds and camera viewpoints. Images are divided into the training split and the validation split at a ratio of about 9:1. **DomainNet** [137] is a recently proposed large-scale dataset which consists of 0.6 million images of 345 classes distributed across 6 domains, *Real*, *Clipart*, *Infograph*, *Painting*, *Quickdraw* and *Sketch*. We follow the training and testing splits for all the 6 domains released by [137]. Also, for the single source DG experiments, we follow [138] and partition the training split from [137] into the training and validation splits at a ratio of 9:1. Due to page limitation, we additionally provide quantitative comparisons on **VLCS** [139] and **Digit-DG** [3] datasets in the supplementary material.

Evaluation Protocol. For fair comparison purposes, we follow the leave-one-domain-out protocol as considered in [3, 112, 116, 113] for our experiments. That is, one data domain from a dataset is selected as the target unseen domain to be recognized, and the remaining ones as the source domains for training. And, we report the top-1 classification accuracy (%) for quantitative evaluation.

doi:10.6342/NTU202301924

Table 3.1: Comparisons to non-data-generation based methods on PACS using ResNet-18 in leave-one-domain-out settings. **Bold** denotes the best result.

Target	DeepAll (baseline)	MMD-AAE [106]	MLDG [107]	JiGen [115]	MetaReg [108]	Epi-FCR [109]	MASF [110]	EISNet [116]	DMG [138]	Borlino *et al.* [143]	DSON [144]	RSC [129]	Ours
Photo	95.6	96.0	96.1	96.0	95.5	93.9	95.0	95.9	93.4	95.0	95.9	96.0	**97.3** ± 0.3
Art painting	75.1	75.2	81.3	79.4	83.7	82.1	80.3	81.9	76.9	82.7	84.7	83.4	**85.8** ± 0.6
Cartoon	74.2	72.7	77.2	75.3	77.2	77.0	77.2	76.4	80.4	78.0	77.7	80.3	**80.7** ± 0.5
Sketch	68.4	64.2	72.3	71.4	70.3	73.0	71.7	74.3	75.2	81.6	**82.2**	80.9	77.3 ± 0.5
Average	78.3	77.0	81.8	80.5	81.7	81.5	81.1	82.2	81.5	84.3	85.1	85.2	**85.3**

3.4.2 Implementation Details

For PACS, Office-Home, and DomainNet, input images are resized to 224 × 224 pixels, and we use ResNet-18 and ResNet-50 [58] pre-trained on ImageNet [140] as the backbones of our teacher and student networks. F_S is trained with the SGD optimizer, with an initial learning rate of 0.0005, and a batch size of 32 for 60 epochs. The learning rate is decayed by 0.1 after 30 epochs. F_T is updated via EMA with the momentum coefficient τ of 0.999 by default. Our novel-domain augmenter G is realized by a fully convolutional network similar to the generator's architecture in [3] and trained with the SGD optimizer. In the warm-up phase, we train F_T together with the classifier C using only source data with the SGD optimizer, and then the parameters of C are fixed in the following training process. Note that we also use the official implementation from [115, 111, 3, 141, 112, 113] for our comparisons. In all our experiments, we implement our model using PyTorch and Dassl.pytorch [142] toolbox, and conduct training on a single NVIDIA TESLA V100 GPU with 32 GB memory.

3.4.3 Quantitative Evaluation

We first perform domain-generalized visual classification tasks and compare our results with existing *non-data-generation* [106, 145, 107, 115, 108, 109, 110, 116, 138, 143, 144, 129] and *data-generation* based [111, 3, 112, 113] methods on two commonly-used public benchmarks, **PACS** and **Office-Home**. In our experiments,

Table 3.2: Comparisons to non-data-generation based methods on Office-Home using ResNet-18 in leave-one-domain-out settings. **Bold** denotes the best result.

Target	DeepAll	CCSA	MMD-	MLDG	D-SAM	JiGen	Borlino	DSON	RSC	Ours
	(baseline)	[145]	AAE [106]	[107]	[146]	[115]	*et al.* [143]	[144]	[129]	
Art	59.0	59.9	56.5	58.1	58.0	53.0	58.7	59.4	58.4	**60.7** ± 0.5
Clipart	48.4	49.9	47.3	49.3	44.4	47.5	52.3	45.7	47.9	**52.9** ± 0.3
Product	72.5	74.1	72.1	72.9	69.2	71.5	73.0	71.8	71.6	**75.8** ± 0.1
Real world	75.5	75.7	74.8	74.7	71.5	72.8	75.0	74.7	74.5	**77.2** ± 0.2
Average	63.9	64.9	62.7	63.8	60.8	61.2	64.8	62.9	63.1	**66.7**

DeepAll is viewed as a baseline, in which both feature extractor and classifier are trained on data aggregated from all source domains.

Tables 3.1 and 3.2 summarize the quantitative comparisons with existing *non-data-generation* based methods [106, 145, 107, 115, 108, 109, 110, 116, 138, 143, 144, 129] on PACS and Office-Home (ResNet-18 as the backbone), respectively. Particularly, Epi-FCR [109] and MASF [110] are meta-learning approaches which either adopt episodic training scheme that cross-train encoders and classifiers from different domains, or employ a gradient-based optimization strategy with global and local losses for regularizing the model training. JiGen [115] and EISNet [116] both consider solving jigsaw puzzles as the auxiliary task for better capturing spatial information. Recent start-of-the-art method RSC [129] iteratively dropouts the most contributing features to force models to explore the remaining features that correlate with semantic information. As we can observe from Table 3.1, our approach achieved the best performance on *Photo*, *Art paining*, and *Cartoon*. It is worth noting that, a significant gap in visual appearance can be seen between *Sketch* and other image domains, which makes the associated domain generalization more difficult. Nevertheless, our approach still achieved satisfactory results over the state-of-the-art methods on *Sketch*, and reported the highest average accuracy of **85.3%**. On the other hand, Table 3.2 demonstrates that our method performed favorably on all the domains (i.e., 60.7% on *Art*, 52.9% on *Clipart*, 75.8% on *Product*, and 77.2% on *Real world*), and thus achieves the highest average accuracy **66.7%**.

Table 3.3: Comparisons to data-generation based methods on PACS using ResNet in leave-one-domain-out settings. **Bold** denotes the best result.

Target	ResNet-18						ResNet-50				
	DeepAll (baseline)	CrossGrad [111]	DDAIG [3]	L2A-OT [112]	MixStyle [113]	Ours	DeepAll (baseline)	CrossGrad [111]	DDAIG [3]	MixStyle [113]	Ours
Photo	95.6	96.0	95.3	96.2	96.1	**97.3** ± 0.3	94.8	97.8	95.7	98.0	**98.9** ± 0.3
Art painting	75.1	79.8	84.2	83.3	84.1	**85.8** ± 0.6	81.5	87.5	85.4	87.4	**90.0** ± 0.3
Cartoon	74.2	76.8	78.1	78.2	78.8	**80.7** ± 0.5	78.6	80.7	78.5	83.3	**83.5** ± 0.5
Sketch	68.4	70.2	74.7	73.6	75.9	**77.3** ± 0.5	69.7	73.9	**80.0**	78.5	**80.0** ± 0.6
Average	78.3	80.7	83.1	82.8	83.7	**85.3**	81.2	85.7	84.9	86.8	**88.1**

The above quantitative comparisons verify that, comparing to directly (meta-)learn from existing source domain data, our approach for augmenting diverse, novel, yet semantically practical source-domain training data would be preferable in domain generalization tasks.

With the above observation, we further compare our method with the state-of-the-art *data-generation* based models [111, 3, 112, 113] using ResNet-18 and ResNet-50 as backbones. As shown in Table 3.3, our approach consistently performed superiorly against the method of [113] by 1.6% and 1.3% on PACS with ResNet-18 and ResNet-50 backbones, respectively. Table 3.4 presents the results on Office-Home, which shows that our method would be preferable among the DG methods considered. Also, the above results demonstrate that our proposed framework is able to achieve general preferable performances regardless of the backbone choice. It is worth noting that, CrossGrad [111] and DDAIG [3] add perturbation to input images, which might not represent the domain variations, and the data generation processes of L2A-OT [112] do *not* jointly take the representation learning into consideration. Also, MixStyle [113] can only produce image features with *interpolated* domain styles. Different from these methods, our approach learns to synthesize out-of-source distribution augmentations and derive domain generalized representations in a mutually beneficial manner, hence exhibiting more robust generalization capability.

Table 3.4: Comparisons to data-generation based methods on Office-Home using ResNet in leave-one-domain-out settings. **Bold** denotes the best result.

Target	ResNet-18						ResNet-50				
	DeepAll	CrossGrad	DDAIG	L2A-OT	MixStyle	Ours	DeepAll	CrossGrad	DDAIG	MixStyle	Ours
	(baseline)	[111]	[3]	[112]	[113]		(baseline)	[111]	[3]	[113]	
Art	59.0	58.4	59.2	60.6	58.7	**60.7** ± 0.5	64.7	67.7	65.2	64.9	**69.3** ± 0.2
Clipart	48.4	49.4	52.3	50.1	**53.4**	52.9 ± 0.3	58.8	57.7	59.2	58.8	**60.1** ± 0.6
Product	72.5	73.9	74.6	74.8	74.2	**75.8** ± 0.1	77.9	79.1	77.7	78.3	**81.5** ± 0.4
Real world	75.5	75.8	76.0	73.0	75.9	**77.2** ± 0.2	79.0	80.4	76.7	78.7	**82.1** ± 0.2
Average	63.9	64.4	65.5	65.6	65.5	**66.7**	70.1	71.2	69.7	70.2	**73.3**

3.4.4 Analysis of Our Method

Ablation Study

We now conduct the ablation study to verify our network design on PACS with ResNet-50 backbone, and we list the results in Table 3.5. Also, we evaluate the effectiveness of *Jigsaw puzzle*. Such spatial transformation has been applied in several DG works [115, 116]. In the bottom part of Table 3.5, we consider different network designs, including *Siamese architecture*, *Student without EMA*, and *Student with EMA*, to be derived for performing on unseen target domains.

Effectiveness of Adversarial Augmenter. In the upper part of this table, we first verify the effectiveness of our designed *novel-domain augmenter* G by replacing G with different types of data augmentation strategies *Random Augmentation* and *Jigsaw puzzle*. *Random Augmentation* denotes directly performing hand-crafted image transformations, including RandomResizedCrop, Color Jittering, Gaussian Blur, RandAugment, and Color Dropping. From Table 3.5, it can be observed that our model surpassed other controlled versions and the baseline on all four domains. We notice that replacing our learnable novel-domain augmenter with hand-crafted random augmentations results in significant performance drops, and the performance was just marginally better than that of the baseline (i.e., *DeepAll*). This verifies that such random image transformations can merely achieve limited improvement on generalization capability. Although the average accuracy of

Jigsaw Puzzle was better than that of *Random Augmentation* by about 2.7%, it was still worse than that of our full version by about 2.6%. This is possibly because that, while *Jigsaw Puzzle* provides more visual clues about spatial information as stated in [115, 116], there is no guarantee that such image transformation would contribute to domain invariance. With the above experiments, our learnable novel-domain augmenter exhibits sufficient ability to generate novel-domain augmentations for facilitating the model robustness to unseen domains.

Effectiveness of Domain Generalized Teacher. From the results shown in the lower half of Table 3.5, we see that the performance dropped when we replace the Teacher-Student scheme by a *Siamese Architecture*, where the parameters are shared between the teacher and student networks. This is due to the fact that the Siamese architecture is prone to output collapsing solutions, hampering the derivation of domain generalized representations. In addition, we examine the performance of applying the trained student network to unseen domains instead of applying Teacher. *Student without EMA* denotes that Teacher is fixed during training, while *Student with EMA* denotes that Teacher is still updated with EMA which benefits the learning of Student. We observe that adopting EMA achieved the better results, but the performance of the above two versions (which apply Student) were still inferior to ours (which applies Teacher). From the above results, we confirm that Teacher updated with EMA would be less likely to be affected by possibly unrealistic domain augmentations during training, avoiding the degradation of its domain generalization capability. As verified by the above experiments, all components presented in our learning scheme would contribute to the domain generalization capability.

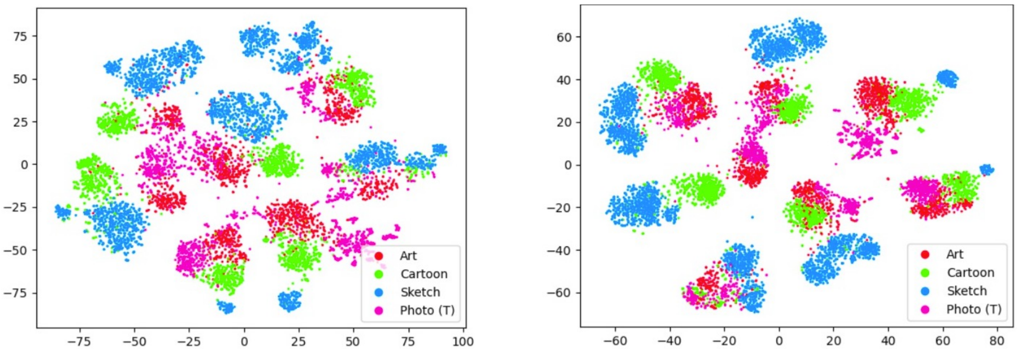
Visualization

We now qualitatively assess the ability of our approach in deriving domain generalizable features. As shown in Fig. 3.2, we apply t-SNE to compare the features z

doi:10.6342/NTU202301924

Table 3.5: Ablation studies on PACS using ResNet-50 as the backbone.

Module	Method	Photo	Art painting	Cartoon	Sketch	Average
Augmentation	DeepAll	94.8	81.5	78.6	69.7	81.2
	Random Aug.	96.4	83.2	75.9	75.5	82.8
	Jigsaw puzzle	97.1	85.3	79.0	**80.5**	85.5
Representation	Siamese archi.	98.3	87.5	82.7	74.5	85.8
	F_S w/o EMA	98.2	86.4	80.1	74.7	84.9
	F_S w/ EMA	97.9	88.9	82.0	75.1	86.0
	Ours $(G + F_T)$	**98.9**	**90.0**	**83.5**	80.0	**88.1**



(a) DeepAll

(b) Our method

Figure 3.2: t-SNE visualization on PACS with Photo as the unseen target domain. (a) Representations extracted by the baseline approach of DeepAll. (b) Representations derived by our approach.

derived by our teacher network F_T with the features extracted by *DeepAll* network on PACS. In this figure, while the source image features extracted by *DeepAll* can be grouped according to their semantic categories, the target-domain features still cannot be properly separated. It can be observed that both source and target-domain features derived by our Teacher are sufficiently aligned, and the distances between different class clusters are more evident, indicating that equipped with our proposed adversarial teacher-student representation learning, our model is capable of learning more discriminative yet domain generalizable features.

Moreover, in Fig. 3.3, we visually compare the synthesized images by our method and those by the state-of-the-art data-generation method of DDAIG [3]

doi:10.6342/NTU202301924

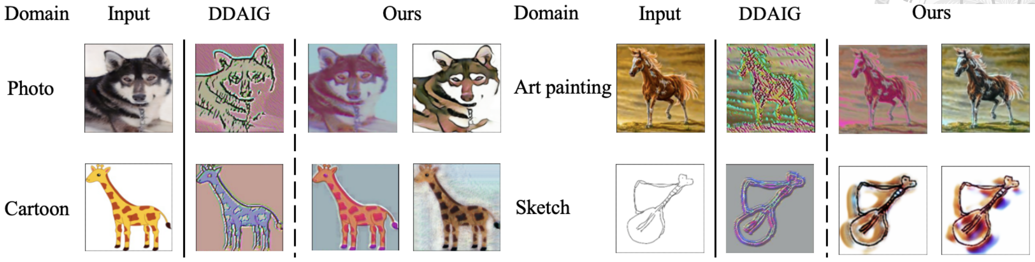


Figure 3.3: Visual comparisons of augmented novel-domain images produced by DDAIG [3] and ours on PACS dataset.

Table 3.6: Impact of momentum coefficient τ on Office-Home using ResNet-50 as the backbone.

τ	Art	Clipart	Product	Real world	Average
0.9	66.5	56.2	78.9	80.9	70.6
0.99	68.1	56.9	80.1	81.4	71.6
0.999	**69.3**	**60.1**	**81.5**	**82.1**	**73.3**

using PACS as the training dataset. As described in Sec. 3.2, [3] learns to perturb the input images for confusing the domain classifier, with the goal of producing output images to be domain-agnostic. However, from Fig. 3.3, we see that images generated by DDAIG [3] tended to exhibit visual perturbation, which might not correspond to domain variations. On the contrary, our approach was capable of producing images in the data domains which are visually realistic yet distinct from source domains. We also note that, our model is trained in a deterministic manner, and the two augmented outputs are generated from our augmenter learned at different time steps with distinct mini-batch data sampled. This supports that our novel-domain augmentation mechanism is able to expand the training distributions.

Impact of the Momentum Coefficient τ

In exponential moving average (EMA), τ is a momentum coefficient to control the update degree of our teacher network F_T. As shown in Table 3.6, we conducted ablation studies on Office-Home with ResNet-50 as the backbone and observed that a large momentum coefficient τ by smoothly refining θ_T could achieve better performance than by rapid updating. These results indicate that a smooth refinement of Teacher avoids the degradation of generalization capability.

3.4.5 Generalization from A Single Source Domain

We evaluate our method on a more challenging DG task, single source domain generalization, to further verify the effectiveness of our method. In the single source DG setting, we only observe training data from a single source domain during training with the aim of generalizing to multiple unseen domains. To confirm that our approach can be extended to the single source DG setting, we conduct experiments on **PACS** and the large-scale benchmark dataset **DomainNet** with the ResNet-50 backbone. For PACS, we select *Photo* as the source domain and the remaining ones (i.e., *Art painting*, *Cartoon*, and *Sketch*) as the target domains. On the other hand, *Real* domain in DomainNet is chosen as the source domain, while *Clipart*, *Infograph*, *Painting*, *Quickdraw*, and *Sketch* domains serve as the target domains. We note that, since only a single source domain is observed during training, the margin m in (3.4) is viewed as a hyperparameter instead of calculating from source domain data. Due to page limitation, additional experiments on PACS using *Art painting*, *Cartoon*, and *Sketch* as the single source domains are presented in the supplementary material.

We provide quantitative comparisons with the baseline (*DeepAll*), JiGen [115], and other three data-generation based methods [111, 3, 141] to evaluate the generalization capability on this challenging setting. As shown in Table 3.7, our approach performed favorably against the baseline (*DeepAll*) and the above DG methods on both benchmark datasets. It is worth noting that, compared with data-generation based methods of [111, 3, 141], our approach was able to achieve superior accuracy on all the target domains of interest. This confirms that, while our method can also be viewed as a data-generation based approach, we are able to better augment novel-domain data based on the observation of single source domain data. From the

doi:10.6342/NTU202301924

Table 3.7: Single-source domain generalization on PACS and DomainNet using ResNet-50 as the backbone. Note that *Photo* of PACS and *Real* of DomainNet are selected as the single source domain for training.

Method	PACS				DomainNet					
	Art painting	Cartoon	Sketch	Average	Clipart	Infograph	Painting	Quickdraw	Sketch	Average
DeepAll	60.7	23.5	29.0	37.7	34.5	15.7	40.7	3.6	25.9	24.1
JiGen [115]	63.6	28.5	30.2	40.8	50.0	19.0	46.3	7.2	35.5	31.6
CrossGrad [111]	64.2	29.4	32.1	41.9	49.4	19.3	47.3	5.8	35.6	31.5
DDAIG [3]	64.1	32.5	29.6	42.1	41.4	16.5	40.9	3.2	26.7	25.7
M-ADA [141]	64.6	34.6	26.6	41.9	50.3	19.5	48.1	7.1	36.0	32.2
Ours	**68.2** ± 0.9	**36.3** ± 0.9	**33.5** ± 0.3	**46.0**	**52.2** ± 0.3	**21.6** ± 0.2	**50.1** ± 0.2	**8.1** ± 0.3	**38.3** ± 0.4	**34.1**

above experiments, the use of our approach for single source domain generalization tasks can be successfully verified.

3.5 Conclusion

In this thesis, we proposed Adversarial Teacher-Student Representation Learning for addressing domain generalization classification tasks. By alternating between the training stages of Teacher-Student representation learning and novel-domain augmentation in an adversarial manner, our learning scheme allows derivation of domain generalizable representations while semantic information is properly preserved. We conduct extensive experiments, including comparisons to state-of-the-art meta-learning and data-generation based DG methods and ablation studies, which quantitatively and qualitatively confirm the effectiveness and robustness of our proposed approach in solving DG classification by training on single or multiple source-domain data.

doi:10.6342/NTU202301924

Chapter 4

Knowledge Transfer for Decentralized Domains

Federated learning (FL) emerges as a decentralized learning framework which trains models from multiple distributed clients without sharing their data to preserve privacy. Recently, large-scale pre-trained models (e.g., Vision Transformer) have shown a strong capability of deriving robust representations. However, the data heterogeneity among clients, the limited computation resources, and the communication bandwidth restrict the deployment of large-scale models in FL frameworks. To leverage robust representations from large-scale models while enabling efficient model personalization for heterogeneous clients, we propose a novel personalized FL framework of client-specific Prompt Generation (pFedPG), which learns to deploy a personalized prompt generator at the server for producing client-specific visual prompts that efficiently adapts frozen backbones to local data distributions. Our proposed framework jointly optimizes the stages of personalized prompt adaptation locally and personalized prompt generation globally. The former aims to train visual prompts that adapt foundation models to each client, while the latter observes local optimization directions to generate personalized prompts for all clients. Through extensive experiments on benchmark datasets, we show that our pFedPG is favorable against state-of-the-art personalized FL methods under

doi:10.6342/NTU202301924

various types of data heterogeneity, allowing computation and communication efficient model personalization.

4.1 Introduction

With access to web-scale training data (*e.g.,* LAION-5B [147]), deep learning has demonstrated remarkable achievements across computer vision [120, 148, 78] and natural language understanding [149, 150, 151]. However, in real-world scenarios, user data is typically scattered across various domains, such as hospital sites or edge devices. Due to increasing risks of privacy breaches and stricter privacy protection regulations [152], centralized learning schemes are not preferable. With the aim of collaboratively training models without exposing users' private data, Federated learning (FL) has emerged as a prominent distributed learning framework and has garnered growing research interest. This privacy-preserving learning paradigm has been widely adopted in applications like medical image diagnosis [153], face recognition [154], and person re-identification [155].

Without the need of data sharing among clients, the mainstream FL approach of FedAvg [156] learns a global model by averaging model parameters trained on clients' private data. However, data distributed in each client might be *heterogeneous* in terms of *domain discrepancy* [157] or *imbalanced class distribution* [158]. Sharing a global model across heterogeneous data clients is prone to highly deviate from their local distribution, leading to severe performance degradation [159, 160]. Previous FL works [161, 158] propose types of constraints (*e.g.,* L_2 [161] or contrastive regularization [158]) to prevent the local training to be divergent from each other. To better handle the inevitable data heterogeneity across clients, personalized federated learning (pFL) methods [159, 160, 162, 163, 164] are instead proposed to allow each client to train a personalized model that adapts to their own data distribution. For example, pFedHN [159] introduces a hypernetwork at the server to directly generate model parameters for each client, whereas pFedLA [160] learns

doi:10.6342/NTU202301924

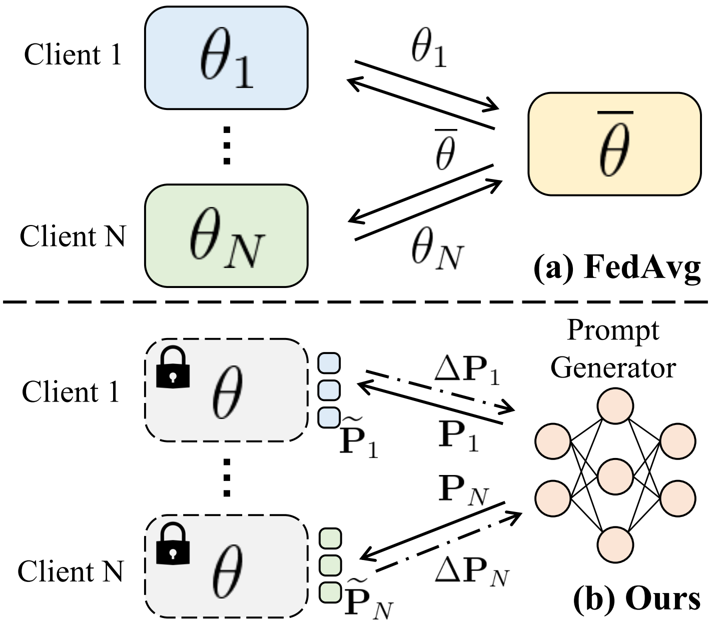


Figure 4.1: Comparison between (a) FedAvg and (b) our approach. Instead of updating and transporting entire models θ, our FL method learns to generate personalized prompts $\mathbf{P}$ by implicitly observing local optimization directions $\Delta\mathbf{P} = \widetilde{\mathbf{P}} - \mathbf{P}$ for efficient model personalization on top of frozen foundation models.

a layer-wise model aggregation policy to assign different weights for personalized model aggregation. While the above pFL approaches are desirable for handling heterogeneous data, they are typically restricted to small backbone architectures (*e.g.,* LeNet [165]) due to the high complexity of outputting model parameters [159] or aggregation weights [160] for large-scale models. Consequently, the capability of derived features is limited, leading to a lack of performance improvement and training instability.

Recently, training from large foundation models [166] for downstream tasks has become a prominent paradigm in centralized learning. To leverage the strong representations derived by foundation models for alleviating data heterogeneity, ViT-FL [167] incorporates pre-trained Vision Transformer (ViT) [168] into standard FL algorithms (*e.g.,* FedAvg [156]) and shows improved robustness and stability on heterogeneously distributed data. However, the use of large pre-trained models for all clients in existing FL algorithms can cause extensive computational and communication burdens, as these methods require transporting entire model parameters between clients and the server. Additionally, overfitting issues might occur when large-scale models are trained with relatively limited client data.

For efficiently tuning large-scale models, prompt learning [169, 170, 171] provides a flexible way to adapt pre-trained models to downstream tasks by solely training the additional inserted trainable parameters (*i.e.,* prompts). For instance, VPT [169] treats prompts as task-specific parameters and prepends them to the input tokens of a pre-trained ViT. In this way, prompts could be optimized to capture task-specific information while instructing a frozen model to perform tasks of interest. However, a straightforward way to adopt prompt learning into FL, *i.e.,* simply averaging prompts learned from all clients, cannot address data heterogeneity among clients effectively and often leads to unsatisfactory performance (as evident in Tables 4.1-4.3). Therefore, there is a crucial challenge to develop new FL methods that can leverage prompt learning effectively while handling data heterogeneity among clients.

In this thesis, we aim at achieving efficient model personalization among clients with data heterogeneity. As depicted in Fig. 4.1, different from conventional FL methods (*e.g.,* FedAvg [156]) that updates and transports entire model parameters, we propose a novel personalized FL scheme of *client-specific Prompt Generation (pFedPG)* that exploits underlying client-specific characteristics to produce personalized prompts for each client, which enables efficient adaptation to local data distribution. To be more precise, each client trains the client-specific prompts to instruct a model to perform recognition tasks on the target client using its private data. As the local training is not required to update entire large models, the computation overload could be minimized while the possible overfitting issues are mitigated accordingly. On the other hand, we employ a personalized prompt generation module on the server side, which is learned to obtain the underlying optimization

doi:10.6342/NTU202301924

directions among clients. With such client characteristics implicitly observed, we are capable of producing personalized prompts to facilitate efficient adaptation for each client with heterogeneous data distribution. By iteratively training the above two stages in a mutually beneficial manner, we are capable of achieving effective yet efficient model personalization on top of the robust representations derived from large-scale foundation models.

We now summarize the contributions of this work below:

- We propose a personalized FL framework of client-specific Prompt Generation (pFedPG), which alternates between *personalized prompt generation* and *personalized prompt adaptation* to enable efficient model personalization under heterogeneous data.

- We design a client-specific prompt generator at the server, which effectively exploits personalized optimization directions and produces client-specific prompts for updating each client model.

- Evaluations on several benchmark datasets in domain discrepancy and imbalanced class distribution verify that our method performs favorably against existing personalized FL approaches and exhibits sufficient training efficiency.

4.2 Related Works

Federated Learning (FL). Federated Learning is a learning framework in machine learning with the goal of training models from distributed data sources while protecting data privacy. The most widely recognized approach for federated learning is FedAvg [156], which partitions the learning process into two phases: local training and global aggregation. Each client first trains its local model using private data and then uploads the model to a server, where the models are averaged to form a global model. However, data distributed in real-world scenarios are

doi:10.6342/NTU202301924

typically non-IID, indicating the presence of domain discrepancy or imbalanced class distribution among clients. Directly averaging models trained on heterogeneous data can lead to severe performance degradation and training instability. To address this challenge, several methods [161, 172, 158, 173, 174, 175, 176] have been proposed to regularize local training in FedAvg [156]. For instance, FedProx [161] and SCAFFOLD [172] restrict the local update to be consistent by L_2 distance over model weights and variance reduction technique over gradients, respectively. MOON [158] applies a contrastive objective to regularize the optimization of local models, ensuring that they do not deviate significantly from the global model. While promising, these methods focus on sharing a global model for all heterogeneous clients, which poses a high risk of deviation from local data distributions.

Recently, personalized Federated Learning (pFL) mechanisms [157, 177, 178, 179, 159, 160, 162, 180, 163, 164] are proposed to address data heterogeneity among clients by learning customized models at each client. Several works [178, 180, 162] achieve model personalization by only aggregating parts of a model (*e.g.,* feature extractor) at the server while keeping or learning additional modules (*e.g.,* classifier) locally. Per-FedAvg [177] analogizes the local training and server aggregation processes as inner and outer loops optimization in model-agnostic meta-learning [181], encouraging local model adaptation from the global model initialization. PartialFed [173] and FedALA [163] derive customized models by adaptively aggregating the global and local models. Similarly, pFedLA [160] learns a layer-wise aggregation policy to construct a personalized model by assigning larger weights to clients with higher similarities. Instead of simply averaging local models at the server, pFedHN [159] directly generates model parameters for all clients in a flexible manner. However, its applicability is limited to small and shallow models (*e.g.,* LeNet [165]) due to the high complexity of the model parameter space.

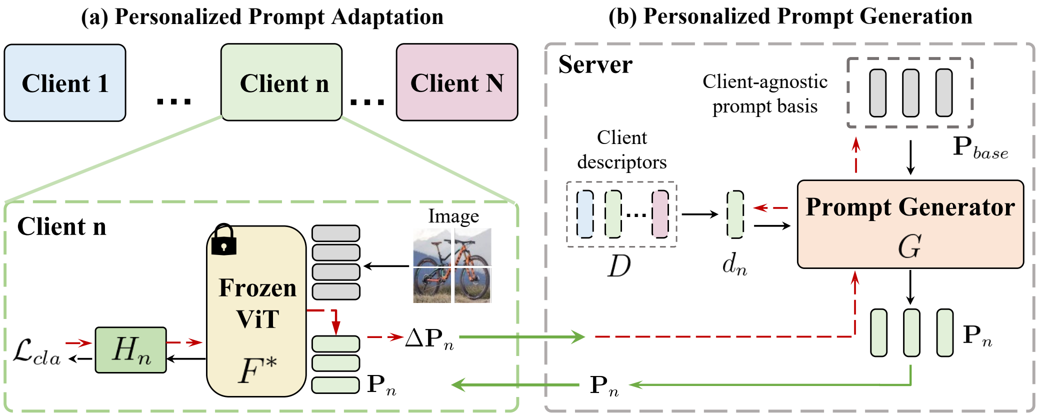


Figure 4.2: Overview of our client-specific Prompt Generation (pFedPG) framework. pFedPG learns a prompt generator G together with client-agnostic prompt basis $\mathbf{P}_{base}$ and a bank of client descriptors $D = \{d_n\}_{n=1}^{N}$ at the server. With local classifcation loss observed, both client-specific prompts $\mathbf{P}_n$ and local classification head H_n are updated at each client n. We alternate between the stages of (a) *personalized prompt adaptation* and (b) *personalized prompt generation* to enable efficient personalization of foundation models like ViT.

Foundation Models and Prompt Learning. Model training from publicly available pre-trained foundation models [166, 168, 120, 148, 78] has emerged as a prominent scheme in centralized learning. In particular, Transformer [182, 168] architectures have demonstrated exceptional ability in deriving robust and discriminative representations. In the FL community, some works [167, 183, 184] start to investigate the effectiveness of initialization and backbone architectures from large foundation models. For instance, ViT-FL [167] first incorporates the pre-trained Vision Transformer (ViT) [168] architecture into FL and shows improved model performance and training stability. However, existing FL algorithms typically require updating the entire model parameters, making the training and transfer of large-scale models challenging in real-world scenarios (*e.g.,* cell phone devices or medical sites), which have limited computation resources or network bandwidth.

Prompt learning techniques [185, 186, 187] have been widely used in the NLP community for adapting language models to downstream tasks effectively via only optimizing a small amount of continuous task-specific prompt vectors. Recently, Visual Prompt Tuning (VPT) [169] has also been proposed as an efficient and effective alternative to full fine-tuning the large-scale ViT model. It introduces additional learnable prompts into the input image embedding space, which serve as task-specific parameters for adapting the frozen backbone to perform a task of interest. However, the integration of prompt learning techniques into heterogeneous federated learning frameworks remains an open research challenge. In this work, we aim to achieve efficient and effective model personalization for heterogeneous data by proposing a personalized prompt generator at the server that implicitly exploits cross-client characteristics to produce personalized prompts for facilitating local adaptation.

4.3 Proposed Method

4.3.1 Problem Formulation

For the sake of completeness, we first define the problem setting in this thesis. Following previous personalized federated learning works [177, 178, 179, 159, 160, 162, 180, 163], we assume that training data are distributed in N separated clients with heterogeneous datasets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$, each contains a set of image-label pairs $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{D}_n|}$. These datasets follow *non-IID* (independent and identically distributed) data distribution in terms of either domain discrepancy or imbalanced label space. With the interest of training efficiency and local data privacy preserved, we aim at learning a client-specific prompt generation mechanism that produces personalized visual prompts $\mathbf{P}_n = [p_n^1, p_n^2, ..., p_n^K]$ that adapt a pre-trained foundation model F^* to perform classification tasks on each local client. Through our learned client-specific prompts, we enable efficient model personalization for each heterogeneous client while preserving the robust representation from a frozen foundation model without the risks of overfitting.

doi:10.6342/NTU202301924

4.3.2 Efficient Model Personalization in FL via Client-Specific Prompt Generation

As illustrated in Fig. 4.2, we propose a personalized federated learning framework of *client-specific Prompt Generation* (pFedPG). To leverage underlying client characteristics and enable efficient model personalization for all clients, pFedPG alternates between the stages of *personalized prompt adaptation* and *personalized prompt generation* in a mutually beneficial manner.

In the stage of *personalized prompt adaptation*, pFedPG advances the visual prompt learning technique [169] in FL frameworks. A small number of trainable parameters, denoted as *prompts* $\mathbf{P}_n = [p_n^1, p_n^2, ..., p_n^K]$, are inserted into a frozen foundation model F^* to encode client-specific information at client n. In the stage of *personalized prompt generation*, a personalized prompt generator G is learned to produce personalized prompts for each client by exploiting the underlying characteristics among clients. Once the learning process is complete, we are able to efficiently adapt the frozen foundation model F^* by the client-specific prompts $\mathbf{P}_n$ to perform recognition tasks at each client n. We now detail each learning stage, including the training/inference processes below.

Personalized prompt adaptation at local clients

To enable efficient model adaptation on top of large-scale foundation models and prevent possible overfitting problems caused by updating on relatively limited private data, we advance *Personalized Prompt Adaptation* based on the prompt learning [169] scheme. Note that, the prompts could be treated as client-specific learnable parameters and directly optimized through gradients during training. With the prompts learned, we can efficiently adapt the foundation model F^* to the data distribution of interest.

As depicted in Fig. 4.2(a), this training stage aims to learn client-specific prompts $\mathbf{P}_n = [p_n^1, p_n^2, ..., p_n^K]$ by leveraging the Transformer-based frozen foundation model F^* with locally updated classification head H_n. To be more specific, we follow [168] and divide an input image $\mathbf{x}$ to m image patches $\{a^i\}_{i=1}^{m}$ and then derive the latent embedding $\mathbf{z}$ by a frozen feature embedding module `Embed` as follows:

$$\begin{aligned}
\mathbf{x} &= [a^1, a^2, ..., a^m], \quad a \in \mathbb{R}^{3\times h\times w}, \\
\mathbf{z} &= [z^1, z^2, ..., z^m], \quad z = \mathtt{Embed}(a),
\end{aligned} \tag{4.1}$$

where h and w denote the height and width of an image patch, and the patch embedding z^m are projected to l-dimension. As the latent embedding $\mathbf{z}$ obtained, we form the input embedding of the Transformer encoder F^* by concatenating $\mathbf{z}$ with a learnable classification token $c \in \mathbb{R}^l$, and the client-specific prompts $\mathbf{P}_n = \left[p_n^1, p_n^2, ..., p_n^K\right]$ as $[c, \mathbf{P}_n, \mathbf{z}]$. To encourage the client-specific prompts to adapt upon this client's data, we employ the standard cross-entropy loss $\mathcal{L}_{cla}$ over $|\mathcal{D}_n|$ samples, and is calculated as:

$$\mathcal{L}_n = \frac{1}{|\mathcal{D}_n|} \sum_{j=1}^{|\mathcal{D}_n|} \mathcal{L}_{cla}\left(H_n\left(F^*\left([c, \mathbf{P}_n, \mathbf{z}_j]\right)\right), y_j\right). \tag{4.2}$$

As a result, the client-specific prompts $\mathbf{P}_\mathtt{n}$ can be optimized end-to-end by gradient decent (the same as H_n and c) with learning rate γ as $\widetilde{\mathbf{P}}_\mathtt{n} \leftarrow \mathbf{P}_\mathtt{n} - \gamma \cdot \partial(\mathcal{L}_n)/\partial\mathbf{P}_\mathtt{n}$.

With the procedure of personalized prompt adaptation training, pFedPG is capable of realizing parameter-efficient model adaptation without requiring updating entire model parameters yet mitigating possible overfitting concerns and huge computation workloads.

Personalized prompt generation at the server

Conventional FL frameworks (*e.g.,* [156]) typically involve the server averaging model parameters from local clients to construct a single global model. However, this aggregation method poses a significant risk of deviating from local data distributions and introduces massive communication overheads, especially when deploying large-scale models among heterogeneous clients. Recall that the prompts trained locally could be treated as client-specific parameters to guide the frozen model in performing recognition tasks for the client of interest. Instead of averaging model parameters or prompts from clients, we aim at learning a unique personalized prompt generation mechanism at the server to exploit cross-client knowledge and then produce personalized prompts that serve as a good initialization for efficient local adaptation at each client. Since the server cannot access local private data, it is challenging to obtain the client-specific characteristics while encouraging the produced personalized prompts to facilitate local adaptation at the same time. In the following, we will elaborate on how our personalized prompt generation be learned in an FL scheme.

Design and architecture. As illustrated in Fig. 4.2(b), with the goal of generating personalized prompts $\{\mathbf{P}_1, ...\mathbf{P}_N\}$ for all N clients, our pFedPG learns to transform a set of client-agnostic prompt basis $\mathbf{P}_{base}$ through a conditional prompt generator $G(\cdot;\varphi)$ parameterized by φ with the guidance of client descriptor d_n selected from $D = \{d_1, d_2, ..., d_N\}$. To be more specific, we realize the conditional prompt generator G based on cross-attention [182] while the client-agnostic prompts $\mathbf{P}_{base}$ and the client descriptor d_n are expected to capture client-agnostic information and encode the client-specific characteristics, respectively. As a result, generating personalized prompts could be achieved by retrieving client-relevant knowledge from $\mathbf{P}_{base}$ through the query of the client descriptor d_n, as formulated below,

$$\begin{aligned}\mathbf{P}_n &= G\left(\mathbf{P}_{base}, d_n\right) = \mathbf{P}_{base} + \texttt{Atten}\left(\mathcal{Q}, \mathcal{K}, \mathcal{V}\right) W^{\mathcal{O}} \\ &= \mathbf{P}_{base} + \texttt{Softmax}(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{l_k}})\mathcal{V}W^{\mathcal{O}},\end{aligned} \tag{4.3}$$

$$\text{where} \quad \mathcal{Q} = [d_n]W^{\mathcal{Q}}, \mathcal{K} = \mathbf{P}_{base}W^{\mathcal{K}}, \mathcal{V} = \mathbf{P}_{base}W^{\mathcal{V}},$$

where $\sqrt{l_k}$ is a scaling factor. $W^{\mathcal{Q}} \in \mathbb{R}^{l \times l_k}$, $W^{\mathcal{K}} \in \mathbb{R}^{l \times l_k}$, $W^{\mathcal{V}} \in \mathbb{R}^{l \times l_v}$, and $W^{\mathcal{O}} \in \mathbb{R}^{l_v \times l}$ are learnable projection metrics as in [182]. As we realize G using a single-head attention layer, the l_k and l_v are equal to the embedding dimension l.

Learning of personalized prompt generation. As the goal of the produced personalized prompts is to serve as a good initialization for each client that facilitates the local adaptation, we learn our personalized prompt generation module (*i.e.*, G, $\mathbf{P}_{base}$ and d_n) through the training rewards observed from the local optimization process. Inspired by [159, 160], the change of prompts after local training $\Delta \mathbf{P}_n = \widetilde{\mathbf{P}_n} - \mathbf{P}_n$ indicates the direction of local optimization at client n that could be treated as a training reward representing the initialization point is good or not. With $\Delta \mathbf{P}_n$ observed, we are capable of training our pFedPG end-to-end via gradient descent.

To be more specific, the update of the conditional prompt generator $G(\cdot;\varphi)$ can be derived by the gradients computed locally and expressed by the chain rule as

$$\begin{aligned} \Delta\varphi = \nabla_{\varphi}\mathcal{L}_n &= (\nabla_{\varphi}\mathbf{P}_n)^T \nabla_{\mathbf{P}_n}\mathcal{L}_n \\ &\cong (\nabla_{\varphi}\mathbf{P}_n)^T \Delta\mathbf{P}_n, \end{aligned} \tag{4.4}$$

where $\nabla_{\mathbf{P}_n}\mathcal{L}_n$ is approximated by $\Delta \mathbf{P}_n$ that indicates the optimization direction of local training. We apply the same optimization rule to learn the client-agnostic prompts $\mathbf{P}_{base}$ and client descriptor d_n end-to-end with G, and summarize the gradient update as follows,

$$\begin{aligned} \varphi &\leftarrow \varphi - \alpha \nabla_{\varphi}\mathbf{P}_n^T \Delta\mathbf{P}_n, \\ \mathbf{P}_{base} &\leftarrow \mathbf{P}_{base} - \alpha \nabla_{\mathbf{P}_{base}}\varphi^T \nabla_{\varphi}\mathbf{P}_n^T \Delta\mathbf{P}_n, \\ d_n &\leftarrow d_n - \alpha \nabla_{d_n}\varphi^T \nabla_{\varphi}\mathbf{P}_n^T \Delta\mathbf{P}_n. \end{aligned} \tag{4.5}$$

We note that, the client-agnostic prompt basis $\mathbf{P}_{base}$ and conditional prompt generator G are optimized by all clients, enforcing them to exploit cross-client knowledge, while client descriptor d_n is solely regarding client n, to encourage the derivation of client-specific characteristics. With our proposed personalized prompt generation module, pFedPG is able to generate personalized prompts to facilitate local adaptation while leveraging and sharing knowledge across clients without explicitly accessing private data.

doi:10.6342/NTU202301924

Algorithm 4: pFedPG for Efficient and Personalized Federated Learning

```
Input: Number of communication rounds $T$, $F^*$, $G$, $\mathbf{P}_{base}$, $D$, and $N$ sets
  of $\mathbf{P}_n$ and $H_n$, $n \in [1, N]$
Data: $N$ labeled datasets $\mathcal{D}_n$, $n \in [1, N]$
Output: $F^*$, $H_n$, $\mathbf{P}_n$
Let $t = 0$;
while $t < T$ do
    # Personalized prompt adaptation at clients
    for $n$ in $1 : N$ do
        Keep $F^*$ freeze;
        Set $\mathbf{P}_n = G(\mathbf{P}_{base}, d_n)$, $d_n \in D$ (Eq. (4.3));
        Randomly sample a minibatch from $\mathcal{D}_n$;
        Update $H_n$ with $\mathcal{L}_n$ (Eq. (4.2));
        Update $\mathbf{P}_n$ by $\widetilde{\mathbf{P}}_{\mathrm{n}} \leftarrow \mathbf{P}_{\mathrm{n}} - \gamma \frac{\partial(\mathcal{L}_n)}{\partial \mathbf{P}_{\mathrm{n}}}$;
        $\Delta \mathbf{P}_n = \widetilde{\mathbf{P}_n} - \mathbf{P}_n$;
    end for
    # Personalized prompt generation at the server
    Receive $\Delta \mathbf{P}_n$ from all $N$ clients;
    Update $G$, $\mathbf{P}_{base}$, and $D$ by Eq. (4.5);
    $t = t + 1$;
end while
```

4.3.3 pFedPG Training and Inference

In Algorithm 4, we summarize the training details of our proposed personalized FL framework of Prompt Generation (pFedPG). We alternate between the learning processes of personalized prompt generation and personalized prompt adaptation until converging.

Once the learning of the proposed framework is complete, we deploy the learned client-specific prompts $\mathbf{P}_n$ to instruct the pre-trained feature extractor F^* to extract

Table 4.1: Quantitative comparisons on Office-Caltech10 and DomainNet datasets. **Bold** denotes the best result.

Datasets	Office-Caltech10 (%)					DomainNet (%)							Comm.
Method	*A*	*C*	*D*	*W*	Avg.	*C*	*I*	*P*	*Q*	*R*	*S*	Avg.	Cost
Baselines													
SingleSet-Full	80.73	73.33	90.62	94.92	84.90	47.34	37.14	67.21	55.30	84.88	45.13	56.17	-
SingleSet-VPT [169]	83.33	74.67	96.88	96.61	87.87	57.98	41.55	74.64	59.60	89.56	60.47	63.97	-
FedAvg [167]	89.58	80.44	100.0	100.0	92.51	63.50	38.05	71.89	60.80	78.55	60.47	62.21	85.88 M
Personalized Federated Learning													
Per-FedAvg [177]	91.67	90.22	100.0	100.0	95.47	69.39	48.71	82.07	35.30	90.63	72.56	66.44	85.88 M
FedRep [178]	91.15	88.44	100.0	100.0	94.90	64.26	38.20	72.86	**62.10**	82.66	60.11	63.37	85.88 M
FedRoD [162]	92.19	90.67	100.0	100.0	95.72	66.54	42.92	74.15	57.20	84.63	66.43	65.31	85.88 M
FedBABU [180]	89.06	85.78	100.0	100.0	93.71	63.31	43.07	74.80	43.80	87.26	67.15	63.23	85.88 M
Efficient Federated Learning													
FedVPT [169]	92.71	84.44	100.0	100.0	94.29	65.59	44.14	76.58	47.30	91.04	60.29	64.16	0.008 M
FedVPT-D [169]	91.67	89.33	100.0	100.0	95.25	63.31	43.07	74.80	54.80	87.26	67.15	65.07	0.009 M
pFedPG (Ours)	**94.79**	**92.44**	**100.0**	**100.0**	**96.81**	**73.00**	**50.08**	**84.33**	60.00	**94.00**	**68.41**	**71.64**	0.008 M

discriminative representations together with locally updated classification head H_n for performing the classification task at each client. Formally, the categorical predictions y^* over Y classes at each client n can be computed as

$$y^* = \arg\min_{k \in K} H_n(F^*([c, \mathbf{P}_n, \mathbf{x}])). \tag{4.6}$$

4.4 Experiments

4.4.1 Datasets and Experimental Setup

Datasets

We evaluate our method on five public benchmark datasets covering types of data heterogeneity, including domain discrepancy and imbalanced class distribution. For *domain discrepancy*, **Office-Caltech10** [188, 189] is composed of four data domains including *Amazon*, *DSLR*, *Webcam*, and *Caltech*. Each domain contains ten classes, with 2,533 images in total. **DomainNet** [137] consists of 0.6 million images of 345 classes distributed across six domains, *Clipart*, *Infograph*, *Painting*, *Quickdraw*, *Real* and *Sketch*. Following [157], we use the top ten most frequent classes to form a sub-dataset for our experiments. As for medical image diagnosis tasks, **Dermoscopic-FL** [153] is comprised of four data sites collected from HAM10K [190] and MSK [191]. Each data site contains three types of skin lesions, with 10,490 images in total. More detailed statistics and sampled images are provided in the supplementary material. For *imbalanced class distribution*, **CIFAR-10** [192] contains 5,000 training images and 1,000 testing images per class, totaling ten classes. **CIFAR-100** [192] consists of 60,000 images of 100 categories with 500 training images and 100 testing images per class. The experimental settings for simulating data heterogeneity are detailed in Sec. 4.4.1.

Experimental settings

To properly evaluate our proposed approach and fairly compare it with existing FL methods, we conduct experiments on two types of heterogeneous FL settings: domain discrepancy and imbalanced class distribution. For conducting clients with *domain discrepancy*, we assign a data domain to a client, indicating the number of clients (N) is set as 4, 6, and 4 for Office-Caltech10, DomainNet, and Dermoscopic-FL datasets, respectively. As for simulating *imbalanced class distribution*, we consider two non-IID settings using CIFAR-10 and CIFAR-100. Following [167], the first non-IID setting we considered is randomly selecting disjoint c classes for each client and denoted as *disjoint label space*. In our experiments, $c = 2$ and $c = 10$ for CIFAR-10 and CIFAR-100, respectively. As for the other non-IID setting, data in each class would be partitioned into all clients following a Dirichlet distribution $Dir(\alpha)$. We follow [162] and set α to 0.1 over 10 clients.

Implementation details

We use ViT-B/16 [168] pre-trained on ImageNet21k [140] as the backbone of F^* and a single linear layer to realize the classification head H_n. The input images of all datasets are resized to 224 × 224 pixels. For each client, we train $\mathbf{P}_{\mathrm{n}}$ and

doi:10.6342/NTU202301924

Table 4.2: Quantitative comparisons on CIFAR-10/100 datasets. **Bold** denotes the best result

Datasets	CIFAR-10 (%)		CIFAR-100 (%)	
Method	Disjoint	$Dir(0.1)$	Disjoint	$Dir(0.1)$
Baselines				
SingleSet-Full	89.51	83.85	67.74	49.64
SingleSet-VPT [169]	88.91	84.32	63.42	46.46
FedAvg [167]	88.04	79.79	63.33	51.37
Personalized Federated Learning				
Per-FedAvg [177]	88.13	85.14	69.31	52.68
FedRep [178]	87.07	82.40	65.71	50.36
FedRoD [162]	87.61	80.36	63.90	51.42
FedBABU [180]	83.15	76.33	55.91	50.19
Efficient Federated Learning				
FedVPT [169]	89.39	85.11	55.49	45.26
FedVPT-D [169]	89.56	85.43	66.91	50.25
pFedPG (Ours)	**90.08**	**87.57**	**70.96**	**55.91**

H_n using the SGD optimizer with a learning rate γ of 0.25 with a weight decay rate of 0.001 and a batch size of 64 for 5 epochs. The number of communication round T is set to 100. We set the learning rate α for updating G, $\mathbf{P}_{base}$, and D to 0.001. The number of prompts K of $\mathbf{P}_{\mathrm{n}}$ and $\mathbf{P}_{base}$ is set as 10 for datasets except for Dermoscopic-FL with $K = 3$. The hyperparameters above are tuned by cross-validation. In all our experiments, we implement our model using PyTorch [193] and conduct training on NVIDIA TESLA V100 GPUs with 32 GB memory.

4.4.2 Quantitative Evaluation

We compare our proposed framework with existing FL methods on benchmark datasets representing various types of data heterogeneity (*i.e.,* domain discrepancy and imbalanced class distribution). In our experiments, *SingleSet-Full* and

FedAvg [156] are viewed as baselines, where the former trains a model at each client without information sharing, while the latter aggregates client models to construct a shared global model. In addition, *SingleSet-VPT* indicates each client independently applies visual prompt tuning [169] to learn prompts at the input embedding space.

In Tables 4.1-4.3, we summarized the results compared with the state-of-the-art pFL works. To be more specific, Per-FedAvg [177] applies meta-learning [181] to derive customized models for each client from a global initialization. FedRep [178] aggregates feature extractors but keeps classifiers trained locally; FedBABU [180] only updates and shares feature extractors during FL training. FedRoD [162] additionally learns a personalized classification head without aggregation to preserve the recognition ability for its own client's data. Instead of updating entire model parameters, two *efficient* FL baselines, *FedVPT* and *FedVPT-D*, are conducted, which keep the backbone frozen, and aggregate prompts globally. Following [169], FedVPT inserts prompts to the input, and FedVPT-D prepends prompts to the input and hidden layers. Note that, we use ViT-B/16 [168] as the backbone of the above methods for fair comparisons.

In Table 4.1, we provide the quantitative comparisons on Office-Caltech10 and DomainNet datasets with the presence of **domain shifts** across clients. Our approach achieved the highest 96.81% and 71.64% average accuracies on Office-Caltech10 and DomainNet, respectively, as observed from Table 4.1. Moreover, our method exhibited the best communication efficiency, accounting for only roughly 0.01% parameters, as compared to the existing pFL methods. In addition to domain discrepancy, we conducted comparisons on the **imbalanced class distribution** scenario using CIFAR-10 and CIFAR-100 datasets, as shown in Table 4.2. As mentioned in Sec. 4.4.1, two types of imbalanced data are simulated, including disjoint label space and imbalanced label distribution drawn from $Dir(0.1)$. Table 4.2 demonstrates that our method performed favorably against existing FL works over the two datasets on both types of label imbalance. To further exhibit the ability

doi:10.6342/NTU202301924

Table 4.3: Quantitative comparisons on Dermoscopic-FL dataset. **Bold** denotes the best result.

Method	A	B	C	D	Avg.
Baselines					
SingleSet-Full	76.09	97.29	71.65	73.57	79.65
SingleSet-VPT [169]	70.90	96.25	70.12	68.33	76.40
FedAvg [167]	62.54	96.12	51.52	68.08	69.57
Personalized Federated Learning					
Per-FedAvg [177]	76.09	91.99	70.12	74.56	78.19
FedRep [178]	69.06	96.12	60.37	68.58	73.53
FedRoD [162]	63.55	96.67	58.84	69.33	72.10
FedBABU [180]	58.19	97.16	49.09	68.58	68.26
Efficient Federated Learning					
FedVPT [169]	74.92	96.77	67.07	75.06	78.46
FedVPT-D [169]	73.91	96.12	74.09	77.81	80.48
pFedPG (Ours)	**79.26**	**97.29**	**76.22**	**78.80**	**82.89**

Table 4.4: Analysis of our personalized prompt generation and the architecture of prompt generator G on benchmark datasets.

Module	Method	Office-Caltech10	DomainNet	CIFAR-10	CIFAR-100
Prompt generation	FedVPT	94.29	64.16	89.39	55.49
	$\mathbf{P}_{base}$	93.16	64.87	88.23	66.89
Architecture of G	MLP [159]	94.96	63.33	87.47	66.73
	AdaIN [6]	95.72	70.08	89.77	69.44
	pFedPG	**96.81**	**71.64**	**90.08**	**70.96**

of our method to more practical and challenging tasks, we provide a comparison with state-of-the-art works for the cross-site medical image diagnosis task using Dermoscopic-FL. As shown in Table 4.3, our pFedPG consistently performed superiorly against other FL methods on all hospital sites.

We observed that, with the presence of significant data heterogeneity (*e.g.*, large style difference in DomainNet) across clients, existing FL works which obtain a shared feature encoder [178, 162, 180] by aggregation might still deviate from local data domains, while Per-FedAvg [177] focuses on deriving a global initialization would not be preferable under severe discrepancy across clients. As shown in Tables 4.1- 4.3, FedVPT and FedVPT-D achieve comparable or even superior performance over existing FL works, exhibiting the ability of efficient FL methods to mitigate possible overfitting issues. However, sharing a set of global prompts is still not desirable for heterogeneous clients. To explicitly enable efficient model personalization to tackle heterogeneous data, our approach learns to generate personalized prompts to facilitate local adaptation for each client. With the above results, we successfully confirm the effectiveness and robustness of our proposed pFedPG to address data heterogeneity with training efficiency.

4.4.3 Analysis of Our pFedPG

In this section, we first conduct experiments to confirm the effectiveness of our proposed personalized prompt generation and the design of our prompt generator G. Second, we provide a detailed analysis of the impact of different number prompts. Due to the page limitations, we provide the analysis of model backbones and the size of client data in the supplementary material.

Effectiveness of personalized prompt generation. In the upper part of Table 4.4, we intend to verify the effectiveness of our personalized prompt generation for facilitating adaptation at each client on benchmark datasets, where CIFAR-10/100 are under the setting of disjoint label space. In Table 4.4, we first ablate $\mathbf{P}_n$ with the global prompts obtained by global averaging (as in *FedVPT*). As reported in Table 4.4, the globally averaged prompts cannot achieve satisfactory performance since sharing a single set of prompts would not be favorable to heterogeneous clients. In addition, we examine the performance of applying the trained *client-agnostic prompt basis* $\mathbf{P}_{base}$ to clients instead of applying personalized prompts $\mathbf{P}_n$.

We observed that the performance of $\mathbf{P}_{base}$ is still inferior to ours (which applies $\mathbf{P}_n$). As evident from the above experiments, the effectiveness of our proposed personalized prompt generation for allowing personalized FL under various types of data heterogeneity would be successfully verified.

Effectiveness of our designed prompt generator G. From the results shown in the lower half of Table 4.4, we see that the performance dropped when we replaced our cross-attention-based prompt generator G and $\mathbf{P}_{base}$ with an MLP-based network as [159], which acts on client descriptors and then output prompts for each client. The inferior performance of the MLP-based prompt generator is due to its high training complexity and instability, resulting from the requirement of deploying a fully-connected layer for each prompt embedding. Another alternative prompt generator is to compute adaptive instance normalization (AdaIN) [6] for $\mathbf{P}_{base}$ and the client descriptor d_n. This method allows for the transfer of client-agnostic prompts $\mathbf{P}_{base}$ to personalized prompts $\mathbf{P}_n$ by replacing the mean and variance calculated from the client descriptor d_n, similar to the style transfer approach [6]. However, as seen in Table 4.4, directly computing AdaIN did not explicitly model the prompt generation process, resulting in inferior performance compared to ours. The results summarized in Table 4.4 confirm the effectiveness of our designed architecture of prompt generator G.

Impact of the number of prompts K. We also analyze the impact of the number of prompts K on benchmark datasets, and show the results in Table 4.5. We found that when the number of prompts is set too low (*e.g.,* $K = 1$), the model's accuracy drops slightly due to insufficient capacity. In contrast, if the number of prompts is set too high, such as 100 or 200, the model's performance significantly degrades. This is because a large number of prompt embeddings may encode noisy and task-irrelevant information, which can adversely affect the quality of the features derived from foundation models. With the above observation, we thus set K as 10 which achieves the best trade-off between communication cost and model accuracy.

doi:10.6342/NTU202301924

Table 4.5: Impact of the number of prompts K on benchmark datasets, where CIFAR-10/100 are drawn from $Dir(0.1)$.

K	Office-Caltech10	DomainNet	CIFAR-10	CIFAR-100
1	96.09	70.27	86.14	55.77
5	96.77	70.53	87.41	55.79
10	**96.81**	**71.64**	**87.57**	**55.91**
50	95.10	69.55	85.63	54.52
100	94.53	68.79	85.02	53.61
200	94.46	66.83	83.53	52.34

4.5 Conclusion

In this thesis, we proposed a novel client-specific Prompt Generation framework (pFedPG) for enabling efficient model personalization among heterogeneous clients. By alternative optimization of the proposed personalized prompt generation and client-specific prompt adaptation, our pFedPG is capable of producing personalized prompts for each client by observing underlying directions of local training among clients, while clients optimize such client-specific prompts to adapt a pre-trained model to local data distribution. We conducted extensive quantitative experiments, verifying that our framework performed favorably against SOTA pFL approaches at heterogeneous data clients while achieving training and communication efficiency.

doi:10.6342/NTU202301924

doi:10.6342/NTU202301924

Chapter 5

Conclusion

In this thesis, we focused on visual understanding for knowledge transfer, addressing the challenges of distribution difference in the aspect of distinct data domains or semantic classes. We systematically examine knowledge transfer for image styles, semantic categories, unseen domains, and decentralized domains. Through our analysis and experimental results, the effectiveness of these approaches is verified.

doi:10.6342/NTU202301924

Reference

[1] A. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, “A unified feature disentangler for multi-domain image translation and manipulation,” *arXiv preprint arXiv:1809.01361*, 2018. viii, 3, 5, 6, 14, 15, 16, 17, 18, 21, 22, 23, 25

[2] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, “Feature generating networks for zero-shot learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551. x, 30, 32, 34, 35, 41, 47, 48, 49, 50, 51, 52, 55, 56

[3] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, “Deep domain-adversarial image generation for domain generalisation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 025–13 032. xi, 60, 63, 70, 71, 73, 74, 76, 77, 78, 79

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” *arXiv preprint arXiv:1508.06576*, 2015. 2

[5] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711. 2

[6] X. Huang and S. J. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization.” in *ICCV*, 2017, pp. 1510–1519. 2, 39, 98, 100

doi:10.6342/NTU202301924

[7] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Advances in Neural Information Processing Systems*, 2017, pp. 386–396. 2

[8] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time hd style transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 698–714. 2

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial nets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 5, 17, 18

[10] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2, 3, 5, 11, 25

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017. 2, 3, 5, 7, 18, 21, 62

[12] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2, 5, 25

[13] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017. 2, 3, 5, 14, 25

[14] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 2, 6, 9, 25, 62

[15] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

doi:10.6342/NTU202301924

[16] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351. 2, 6, 25, 62

[17] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2765–2773. 2, 25

[18] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017. 2, 3, 6, 7, 25

[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2, 4, 12

[20] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2, 4

[21] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 2, 4, 10

[22] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. F. Wang, "Detach and adapt: Learning cross-domain disentangled deep representation," *arXiv preprint arXiv:1705.01314*, 2017. 2, 3, 5, 6, 14, 18, 25

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014. 3, 32, 34, 67

doi:10.6342/NTU202301924

[24] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *arXiv preprint arXiv:1804.04732*, 2018. 3, 5, 14, 21, 22, 23

[25] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," *arXiv preprint arXiv:1808.00948*, 2018. 3, 5, 11, 14, 16, 17, 21, 22, 23

[26] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797. 3

[27] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, 2017, pp. 465–476. 3, 5, 18

[28] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Advances in Neural Information Processing Systems*, 2018, pp. 1287–1298. 3, 5

[29] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," *arXiv preprint arXiv:1707.04993*, 2017. 4

[30] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424. 4, 5

[31] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Multi-task adversarial network for disentangled feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3743–3751. 4, 5

doi:10.6342/NTU202301924

[32] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1623–1632. 4

[33] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "Cr-gan: learning complete representations for multi-view generation," *arXiv preprint arXiv:1806.11191*, 2018. 4, 5

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 4, 34, 35

[35] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 5

[36] Z. Yi, H. Zhang, P. T. Gong *et al.*, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[37] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *arXiv preprint*, vol. 1711, 2017. 5, 6

[38] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *arXiv preprint arXiv:1711.10678*, 2017. 5, 6

[39] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 37–52. 6

doi:10.6342/NTU202301924

[40] Z. Ding and Y. Fu, "Deep transfer low-rank coding for cross-domain learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 6, pp. 1768–1779, 2018. 6

[41] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. 17

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 17

[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 18

[44] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 783–790. 18

[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637. 22

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595. 22

[47] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 25

doi:10.6342/NTU202301924

[48] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, “Deep reconstruction-classification networks for unsupervised domain adaptation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 597–613. 25

[49] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 4. 25

[50] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, “Generate to adapt: Aligning domains using generative adversarial networks,” *arXiv preprint arXiv:1704.01705*, 2017. 25

[51] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, “From source to target and back: symmetric bi-directional adaptive gan,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8099–8108. 25

[52] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, “Evaluation of output embeddings for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936. 30, 31, 33, 48, 49

[53] L. Zhang, T. Xiang, and S. Gong, “Learning a deep embedding model for zero-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2021–2030. 30, 31, 34

[54] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, “Generalized zero-and few-shot learning via aligned variational autoencoders,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8247–8255. 30, 34, 49, 51

[55] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, “f-vaegan-d2: A feature generating framework for any-shot learning,” in *Proceedings of the IEEE*

doi:10.6342/NTU202301924

Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 275–10 284. 30, 32, 34, 35, 47, 48, 49, 51

[56] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging seen and unseen semantic relationships for generative zero-shot learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 70–86. 30, 34, 35, 47, 49, 51, 52

[57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9. 29

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 29, 36, 71

[59] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448. 29

[60] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 29

[61] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 30

[62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969. 30

[63] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE transactions on*

doi:10.6342/NTU202301924

pattern analysis and machine intelligence, vol. 36, no. 3, pp. 453–465, 2013. 31

[64] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018. 31, 50

[65] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *European Conference on Computer Vision*. Springer, 2016, pp. 52–68. 31, 32

[66] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, and X.-Z. Wang, "A review of generalized zero-shot learning methods," *arXiv preprint arXiv:2011.08641*, 2020. 31

[67] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015. 31, 33, 48, 49

[68] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3174–3183. 31, 33, 49

[69] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," *arXiv preprint arXiv:1412.6568*, 2014. 31, 34

[70] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto, "Ridge regression, hubness, and zero-shot learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 135–151. 31, 34

doi:10.6342/NTU202301924

[71] S. Changpinyo, W.-L. Chao, and F. Sha, “Predicting visual exemplars of unseen classes for zero-shot learning,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3476–3485. 31

[72] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro, “Multi-modal cycle-consistent generalized zero-shot learning,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 21–37. 32, 34, 35, 47, 48, 49, 51

[73] Y. Zhu, J. Xie, B. Liu, and A. Elgammal, “Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9844–9854. 32, 34, 47, 48, 49

[74] J. Li, M. Jing, K. Lu, L. Zhu, Y. Yang, and Z. Huang, “Alleviating feature confusion for generative zero-shot learning,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1587–1595. 32

[75] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, “Devise: A deep visual-semantic embedding model,” in *Advances in neural information processing systems*, 2013, pp. 2121–2129. 33, 48, 49

[76] B. Romera-Paredes and P. Torr, “An embarrassingly simple approach to zero-shot learning,” in *International Conference on Machine Learning*, 2015, pp. 2152–2161. 33, 48, 49

[77] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” in *Advances in neural information processing systems*, 2013, pp. 935–943. 33, 49

[78] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual

doi:10.6342/NTU202301924

models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. 34, 82, 87

[79] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7402–7411. 34, 47, 48, 49, 51

[80] J. Ni, S. Zhang, and H. Xie, "Dual adversarial semantics-consistent network for generalized zero-shot learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 6143–6154. 34, 35, 49, 51

[81] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2188–2196. 34

[82] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017. 35, 43

[83] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3018–3027. 35, 55

[84] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, and A. Bronstein, "Delta-encoder: an effective sample synthesis method for few-shot object recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 2845–2855. 35

[85] M. Chen, Y. Fang, X. Wang, H. Luo, Y. Geng, X. Zhang, C. Huang, W. Liu, and B. Wang, "Diversity transfer network for few-shot learning." in *AAAI*, 2020, pp. 10 559–10 566. 35

doi:10.6342/NTU202301924

[86] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, “Mode seeking generative adversarial networks for diverse image synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1429–1437. 36

[87] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, “Drit++: Diverse image-to-image translation via disentangled representations,” *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020. 36

[88] B. AlBahar and J.-B. Huang, “Guided image-to-image translation with bi-directional feature transformation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9016–9025. 39

[89] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, “Few-shot unsupervised image-to-image translation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 551–10 560. 39

[90] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, “Caltech-ucsd birds 200,” 2010. 46

[91] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 951–958. 46

[92] G. Patterson and J. Hays, “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2751–2758. 46

[93] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729. 47

doi:10.6342/NTU202301924

[94] S. Reed, Z. Akata, H. Lee, and B. Schiele, “Learning deep representations of fine-grained visual descriptions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58. 47

[95] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208. 48, 49

[96] X. Wang, F. Yu, R. Wang, T. Darrell, and J. E. Gonzalez, “Tafe-net: Task-aware feature embeddings for low shot learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1831–1840. 47, 48, 49, 51

[97] H. Jiang, R. Wang, S. Shan, and X. Chen, “Transferable contrastive network for generalized zero-shot learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9765–9774. 47, 48, 49, 51

[98] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent embeddings for zero-shot classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 69–77. 47, 48, 49

[99] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal, “A generative adversarial approach for zero-shot learning from noisy texts,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1004–1013. 47, 48, 49

[100] R. Gao, X. Hou, J. Qin, L. Liu, F. Zhu, and Z. Zhang, “A joint generative model for zero-shot learning,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0. 48, 49, 51

[101] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, “Generalized zero-shot learning via synthesized examples,” in *Proceedings of the IEEE conference*

doi:10.6342/NTU202301924

on computer vision and pattern recognition, 2018, pp. 4281–4289. 47, 48, 49

[102] Y. Liu, J. Guo, D. Cai, and X. He, “Attribute attention for semantic disambiguation in zero-shot learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6698–6707. 49

[103] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, “Synthesized classifiers for zero-shot learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5327–5336. 49

[104] H. Yu and B. Lee, “Zero-shot learning via simultaneous generating and learning,” in *Advances in Neural Information Processing Systems*, 2019, pp. 46–56. 49

[105] R. Keshari, R. Singh, and M. Vatsa, “Generalized zero-shot learning via over-complete distribution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 300–13 308. 47, 49, 51

[106] H. Li, S. J. Pan, S. Wang, and A. C. Kot, “Domain generalization with adversarial feature learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5400–5409. 60, 71, 72

[107] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, “Learning to generalize: Meta-learning for domain generalization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. 60, 63, 71, 72

[108] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, “Metareg: Towards domain generalization using meta-regularization,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 998–1008, 2018. 60, 63, 66, 71, 72

doi:10.6342/NTU202301924

[109] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1446–1455. 60, 63, 66, 71, 72

[110] Q. Dou, D. C. Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *arXiv preprint arXiv:1910.13580*, 2019. 60, 63, 66, 71, 72

[111] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," *arXiv preprint arXiv:1804.10745*, 2018. 60, 63, 71, 73, 74, 78, 79

[112] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 561–578. 60, 63, 70, 71, 73, 74

[113] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," *arXiv preprint arXiv:2104.02008*, 2021. 60, 63, 70, 71, 73, 74

[114] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017. 60

[115] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238. 61, 63, 64, 71, 72, 74, 75, 78, 79

[116] S. Wang, L. Yu, C. Li, C.-W. Fu, and P.-A. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 159–176. 61, 64, 70, 71, 72, 74, 75

doi:10.6342/NTU202301924

[117] S. Bucci, A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-supervised learning across domains," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 61

[118] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717. 61, 64

[119] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 61, 64

[120] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020. 61, 64, 67, 82, 87

[121] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020. 61, 64

[122] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning," *arXiv preprint arXiv:2005.10243*, 2020. 61, 64, 67

[123] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised learning from a multi-view perspective," *arXiv preprint arXiv:2006.05576*, 2020. 61, 64, 67

[124] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017. 61, 67

[125] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, “Unbiased teacher for semi-supervised object detection,” *arXiv preprint arXiv:2102.09480*, 2021. 61, 67

[126] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998. 62

[127] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, “Crdoco: Pixel-level domain transfer with cross-domain consistency,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1791–1800. 62

[128] Y. Li, Y. Yang, W. Zhou, and T. M. Hospedales, “Feature-critic networks for heterogeneous domain generalization,” *arXiv preprint arXiv:1901.11448*, 2019. 63

[129] Z. Huang, H. Wang, E. P. Xing, and D. Huang, “Self-challenging improves cross-domain generalization,” *arXiv preprint arXiv:2007.02454*, vol. 2, 2020. 63, 71, 72

[130] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” *arXiv preprint arXiv:2103.02503*, 2021. 63

[131] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019. 63

[132] L. Li, K. Gao, J. Cao, Z. Huang, Y. Weng, X. Mi, Z. Yu, X. Li, and B. Xia, “Progressive domain expansion network for single domain generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 224–233. 63

doi:10.6342/NTU202301924

[133] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European conference on computer vision*. Springer, 2016, pp. 577–593. 64

[134] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018. 64

[135] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550. 70

[136] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5018–5027. 70

[137] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019. 70, 94

[138] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to balance specificity and invariance for in and out of domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 301–318. 70, 71, 72

[139] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1657–1664. 70

[140] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009. 71, 95

doi:10.6342/NTU202301924

[141] F. Qiao, L. Zhao, and X. Peng, “Learning to learn single domain generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 556–12 565. 71, 78, 79

[142] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, “Domain adaptive ensemble learning,” *arXiv preprint arXiv:2003.07325*, 2020. 71

[143] F. C. Borlino, A. D’Innocente, and T. Tommasi, “Rethinking domain generalization baselines,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9227–9233. 71, 72

[144] S. Seo, Y. Suh, D. Kim, J. Han, and B. Han, “Learning to optimize domain specific normalization for domain generalization,” *arXiv preprint arXiv:1907.04275*, vol. 3, no. 6, p. 7, 2019. 71, 72

[145] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, “Unified deep supervised domain adaptation and generalization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725. 71, 72

[146] A. D’Innocente and B. Caputo, “Domain generalization with domain-specific aggregation modules,” in *German Conference on Pattern Recognition*. Springer, 2018, pp. 187–198. 72

[147] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *arXiv preprint arXiv:2210.08402*, 2022. 82

[148] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *CVPR*, 2022. 82, 87

doi:10.6342/NTU202301924

[149] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 82

[150] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022. 82

[151] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020. 82

[152] B. Custers, A. M. Sears, F. Dechesne, I. Georgieva, T. Tani, and S. Van der Hof, *EU personal data protection in policy and practice*. Springer, 2019. 82

[153] Z. Chen, M. Zhu, C. Yang, and Y. Yuan, "Personalized retrogress-resilient framework for real-world medical federated learning," in *MICCAI*, 2021. 82, 95

[154] C.-T. Liu, C.-Y. Wang, S.-Y. Chien, and S.-H. Lai, "Fedfr: Joint optimization federated framework for generic and personalized face recognition," in *AAAI*, 2022. 82

[155] W. Zhuang, Y. Wen, X. Zhang, X. Gan, D. Yin, D. Zhou, S. Zhang, and S. Yi, "Performance optimization of federated person re-identification via benchmark analysis," in *ACM MM*, 2020. 82

[156] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017. 82, 83, 84, 85, 86, 90, 97

doi:10.6342/NTU202301924

[157] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, “Fed{bn}: Federated learning on non-{iid} features via local batch normalization,” in *ICLR*, 2021. 82, 86, 95

[158] Q. Li, B. He, and D. Song, “Model-contrastive federated learning,” in *CVPR*, 2021. 82, 86

[159] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, “Personalized federated learning using hypernetworks,” in *ICML*, 2021. 82, 83, 86, 88, 92, 98, 100

[160] X. Ma, J. Zhang, S. Guo, and W. Xu, “Layer-wised model aggregation for personalized federated learning,” in *CVPR*, 2022. 82, 83, 86, 88, 92

[161] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in *MLSys*, 2020. 82, 86

[162] H.-Y. Chen and W.-L. Chao, “On bridging generic and personalized federated learning for image classification,” in *ICLR*, 2022. 82, 86, 88, 94, 95, 96, 97, 98, 99

[163] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, “Fedala: Adaptive local aggregation for personalized federated learning,” *arXiv preprint arXiv:2212.01197*, 2022. 82, 86, 88

[164] Y. Shen, Y. Zhou, and L. Yu, “Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning,” in *CVPR*, 2022. 82, 86

[165] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, 1998. 83, 86

[166] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, “On the opportu-

doi:10.6342/NTU202301924

nities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. 83, 87

[167] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, and D. Rubin, "Rethinking architecture design for tackling data heterogeneity in federated learning," in *CVPR*, 2022. 83, 87, 94, 95, 96, 98

[168] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021. 83, 87, 90, 95, 97

[169] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *ECCV*, 2022. 84, 88, 89, 94, 96, 97, 98

[170] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 84

[171] ——, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, 2022. 84

[172] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *ICML*, 2020. 86

[173] B. Sun, H. Huo, Y. Yang, and B. Bai, "Partialfed: Cross-domain personalized federated learning via partial initialization," *NeurIPS*, 2021. 86

[174] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," in *AAAI*, 2022. 86

doi:10.6342/NTU202301924

[175] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, “Fine-tuning global model via data-free knowledge distillation for non-iid federated learning,” in *CVPR*, 2022. 86

[176] M. Mendieta, T. Yang, P. Wang, M. Lee, Z. Ding, and C. Chen, “Local learning matters: Rethinking data heterogeneity in federated learning,” in *CVPR*, 2022. 86

[177] A. Fallah, A. Mokhtari, and A. Ozdaglar, “Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach,” in *NeurIPS*, 2020. 86, 88, 94, 96, 97, 98, 99

[178] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, “Exploiting shared representations for personalized federated learning,” in *ICML*, 2021. 86, 88, 94, 96, 97, 98, 99

[179] T. Li, S. Hu, A. Beirami, and V. Smith, “Ditto: Fair and robust federated learning through personalization,” in *ICML*, 2021. 86, 88

[180] J. Oh, S. Kim, and S.-Y. Yun, “Fedbabu: Toward enhanced representation for federated image classification,” in *ICLR*, 2022. 86, 88, 94, 96, 97, 98, 99

[181] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*, 2017. 86, 97

[182] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017. 87, 91

[183] J. Nguyen, J. Wang, K. Malik, M. Sanjabi, and M. Rabbat, “Where to begin? on the impact of pre-training and initialization in federated learning,” *arXiv preprint arXiv:2210.08090*, 2022. 87

doi:10.6342/NTU202301924

[184] H.-Y. Chen, C.-H. Tu, Z. Li, H.-W. Shen, and W.-L. Chao, “On the importance and applicability of pre-training for federated learning,” in *ICLR*, 2023. 87

[185] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” *arXiv preprint arXiv:2101.00190*, 2021. 87

[186] X. Liu, K. Ji, Y. Fu, Z. Du, Z. Yang, and J. Tang, “P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks,” *arXiv preprint arXiv:2110.07602*, 2021. 87

[187] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” *arXiv preprint arXiv:2104.08691*, 2021. 87

[188] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in *ECCV*, 2010. 94

[189] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” Tech. Rep., 2007. 94

[190] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific data*, 2018. 95

[191] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *ISBI*, 2018. 95

[192] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009. 95

doi:10.6342/NTU202301924

[193] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *NeurIPS*, 2019. 96

doi:10.6342/NTU202301924