Learning Classifiers for Target Domain with Limited or No Labels

Pengkai Zhu [*1] **Hanxiao Wang** [*1] **Venkatesh Saligrama** [1]

Abstract

In computer vision applications, such as domain adaptation (DA), few shot learning (FSL) and zero-shot learning (ZSL), we encounter new objects and environments, for which insufficient examples exist to allow for training "models from scratch," and methods that adapt existing models, trained on the presented training environment, to the new scenario are required. We propose a novel visual attribute encoding method that encodes each image as a low-dimensional probability vector composed of prototypical part-type probabilities. The prototypes are learnt to be representative of all training data. At test-time we utilize this encoding as an input to a classifier. At test-time we freeze the encoder and only learn/adapt the classifier component to limited annotated labels in FSL; new semantic attributes in ZSL. We conduct extensive experiments on benchmark datasets. Our method outperforms state-of-art methods trained for the specific contexts (ZSL, FSL, DA).

1. Introduction

Deep Neural Networks have emerged as the state-of-art in terms of achievable accuracy in a wide variety of applications (Zhu et al., 2019a; Chen et al., 2019; Zhang et al., 2018b) including large-scale visual classification. Nevertheless, this success of DNNs has critically hinged on availability of large amount of labeled training data, data that is labeled by human labelers. It is increasingly being recognized, particularly in the context of large-scale visual classification problems (Russakovsky et al., 2014), that such large-scale human labeling is not scalable (Antol et al., 2014), and we must account for challenges posed by non-uniform and sparsely annotated training data (Bhatia et al., 2015) as in few-shot learning (FSL), appearance of novel objects for in-the-wild scenarios (Zhu et al., 2019b) as in generalized zero-shot learning (GZSL), and responding to changes in operational environment such as changes in data collection viewpoints as exemplified by domain adaptation (DA) (Tzeng et al., 2017).

Decomposability and Compositionality: We are motivated to respond to these aforementioned challenges without training "models from scratch," which requires collecting new labeled data, and yet achieving high-accuracy. We propose a novel framework and DNN architecture that addresses these challenges in a unified manner. Our key insight is based on decomposability of objects into proto-typical primitive parts/part-types and compositionality of proto-typical primitive part/part-types to explain new, unseen or modified object classes. This insight is not new and has been employed in a long-line of work particularly in cognitive science to explain human concept learning such as children making meaningful generalization through one-shot learning, parsing objects into parts, and generating new concepts from parts (see (Lake et al., 2015)). While (Lake et al., 2015) advocates a generative Bayesian Program Learning framework to mimic human concept learning and avoid "data-hungry" DNNs altogether, we advocate use of DNNs and situate our work within a discriminative learning framework and employ novel DNN architectures that also obviates the need for new annotated data and realizes high-accuracy.

Our Contributions. We propose a novel approach that encodes an input instance as a collection of probability vectors. Each probability vector is associated with a part and represents the mixture of prototypical part types that makeup the part. To do this we train a Multi-Attention CNN (MACNN), which produces a diverse collection of attention regions and associated features masking out uninteresting regions of the image space. These attention regions are decomposed into a suitably small number of prototypical parts and prototypical part probabilities, yielding a low-dimensional encoding. We refer to these encodings as low-dimensional visual attribute (LDVA) encodings since they are analogous to how humans would quantify the existence of an attribute in the presented instance by drawing similarity to a prototypical attribute seen from experience. We input the LDVA encoding into a predictor, which then predicts the output for the different scenarios (ZSL, FSL or DA). We learn an end-to-end model on training data and at test-time, freeze the high-dimensional mapping to LDVA encoding component, and only adapting the predictor based on what is revealed during test-time.

[*]Equal contribution [1]Electrical and Computer Engineering Department, Boston University, Boston, USA. Correspondence to: Venkatesh Saligrama <srv@bu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Training & Test-time Prediction: For unsupervised domain adaptation (UDA), both annotated source data and unannotated target data are utilized for training our end-to-end model. However, since no additional data is available, both LDVA and the classifier are unchanged at test-time. For GZSL, we assume a one-to-one correspondence between semantic vectors and class labels as is the convention. During training we assume access to seen class image instances and associated semantic vectors, while being agnostic to both unseen images and unseen semantic vectors. At test-time, we fix our LDVA embedding and modify the prediction component to incorporate semantic vectors from all seen and unseen classes. Finally, for FSL, we learn only the classifier using LDVA as inputs.

Why is LDVA effective? Our results on benchmark datasets highlights the utility of mapping visual instances into the LDVA encoding and its tolerance to visual distortions. For DA, while an image can exhibit significant visual distortion and so domain shifts, the LDVA encodings for source and target are similar[1], thus obviating the need to modify the classifier (see *Figure 3*). For GZSL, the LDVA encoding mirrors how semantic attributes are scored. This enables meaningful knowledge transfer from visual to semantic domain and reducing the semantic gap (see *Figure 2*).

Section 2 describes related work. In Section 3 we first present an overview of proposed approach and then later describe concretely our models. In Section 4 we describe experiments on benchmark datasets for DA, FSL and ZSL.

2. Related Work

Related approaches for adapting models from presented training environment (PTE) can be divided into three groups: pixel-space methods, feature-space methods, and latent-space methods. In contrast to our LDVA encoding that encodes the mixture composition of parts, these works typically attempt to transfer knowledge in a high-dimensional space. We list different lines of research in this context.

Pixel-space methods focus on generating or synthesizing images in the new visual environments, or converting new environment images into existing PTE, so as to avoid exhaustive human annotation. These works are largely based on the recently proposed Generative Adversarial Networks (Goodfellow et al., 2014). In domain adaptation, (Taigman et al., 2016; Shrivastava et al., 2017; Bousmalis et al., 2017) propose to train a generator to transform a source image into a target image (or vice versa) and meanwhile force the generated image to be similar to the original one. (Liu & Tuzel, 2016) trains a tuple of GANs for both domains and ties the weights for certain layers to jointly learn a source and target representation. (Ghifary et al., 2016) enforces the features learnt on the source data to reconstruct target images to encourage alignment in the unsupervised adaptation setting. In generalized zero-shot learning, analogous to domain adaptation, attempts have been made on synthesizing unseen class images in the new environment from the given semantic attributes, e.g. (Zhu et al., 2018; Kumar Verma et al., 2018; Xian et al., 2018b; Jiang et al., 2018). In few-shot learning, generative models are often used for data augmentation to account for the sparsely labelled few-shot examples, e.g. (Antoniou et al., 2017; Wang et al., 2018c; Mehrotra & Dukkipati, 2017).

Feature-space methods propose to either directly learn environment-invariant feature/predictor model, or align the models from the new and source environments to address the problem of insufficient annotations. For instance, several domain adaptation works propose learning a domain-invariant feature embedding via adversarial training (Long et al., 2018; Tzeng et al., 2017) and graph-based label propagation (Ding et al., 2018), while others propose aligning the target domain feature distribution to source domain, e.g. (Kumar et al., 2018). There are also methods which perform adaptation in both feature-space and pixel-space. For example, (Hoffman et al., 2017) proposes a model which adapts between domains using both generative image space alignment and latent representation space alignment. Similar approaches have also been investigated in GZSL. (Frome et al., 2013; Zhang & Saligrama, 2016; Lee et al., 2018; Wang et al., 2018b) propose learning feature embeddings that directly map the visual domain to the semantic domain and infer classifiers for unseen classes. In (Annadani & Biswas, 2018; Kodirov et al., 2017), authors propose an encoder-decoder network with the goal of mirroring learnt semantic relations between different classes in the visual domain. In FSL, (Sung et al., 2018) and (Vinyals et al., 2016) propose adopting an environment-invariant feature representation that is based on comparing an input sample to a support set and use the similarity scores for classification input in the new environment.

Latent-space methods aim to discover latent feature spaces for PTE images that are universal and agnostic to environment changes, and thus can be further used as a general representation for newly encountered images in a new environment. For example, these latent variables include locations of the attention regions on interesting foreground objects, clusters and manifolds information of the data distributions, or common visual part features (which are still high-dimensional). For DA, (Kang et al., 2018) assumes attention of the convolutional layers to be invariant to the domain shift and propose aligning the attentions for source and target domain images. (Wang et al., 2018a) learns Grassmann manifold with structural risk minimization, and train a domain-invariant classifier on the learnt manifold. (Shu

[1] consider handwritten digits under going a domain shift but the composition of parts that make up the digit is quite similar

et al., 2018) makes use of the data distribution by first clustering the data and assumes samples in the same cluster share the same label. Target domain is modified so as to not break clusters. In GZSL, (Li et al., 2018) propose zoom-net as a means to filter-out redundant visual features such as deleting background and focus attention on important locations of an object. (Zhu et al., 2018) further extend this insight and propose visual part detector (VPDE-Net) and utilize high-dimensional part feature vectors as an input for semantic transfer, namely, to synthesize unseen examples by leveraging knowledge of unseen class attributes. Similarly in FSL, (Snell et al., 2017) propose learning prototypical representations of each class by hard clustering on a support set and perform classification on these representations. (Lin et al., 2017) finds that training manifolds in 3D views results in manifolds that are more general and abstract, likely at the levels of parts, and independent of the specific objects or categories in the PTE. There are also the family of meta-learning methods, e.g. (Ravi & Larochelle, 2016; Munkhdalai & Yu, 2017; Finn et al., 2017) which treats the model/optimization parameters as latent variables and propose meta-models to infer such parameters.

3. Methodology

3.1. Problem Definition

The problem scenarios under consideration consist of source and target domains and call for prediction on target domain by means of training data available in different forms. We denote by $x \in \mathcal{X} \subset \mathbb{R}^D$ inputs taking values in a feature space and $y \in \mathcal{Y}$ the output labels taking values in a finite set $\mathcal{Y}$ and $p(x, y)$ the joint distribution. Whenever necessary, we denote by $p_s(x, y)$, $p_t(x, y)$ source and target joint distributions respectively. Since we focus primarily on images of fixed dimension, we assume that the input space for source and target domain is the same. We allow for class labels for source and target domains to be different. We use superscript notation: $y^s, \mathcal{Y}^s, y^t, \mathcal{Y}^t$ when necessary for source and target labels to avoid confusion.

Unsupervised Domain Adaptation (UDA): Source and target domain spaces share same labels. The joint distributions $p_s(y|x) \neq p_t(y|x)$ and $p_s(x|y) \neq p_t(x|y)$. For training, we are provided n_s IID instances of annotated source domain data $(x_i, y_i) \overset{d}{\sim} p_s(x, y),\ i \in [n_s]$ and n_t IID instances of unannotated input instances $x_i \overset{d}{\sim} p_t(x) = \sum_{y \in \mathcal{Y}} p_t(x, y)$. Our goal is to learn a predictor $f(\cdot)$ that generalizes well, i.e., the expected loss $\bar{L}_t = \mathbb{E}_{(x,y)\sim p_t} \mathbb{1}_{\{f(x) \neq y\}}$ is small.

Few Shot Learning (FSL): Note that while UDA and FSL share some similarities, i.e., $p_s(x, y) \neq p_t(x, y)$, they are different cases because, in FSL, the collection of source and target labels are not identical and could even be mutually exclusive. In FSL, we have two datasets during the training stage, i.e. a training set and a support set. The training set contains data from several source domains, $s_j,\ j \in [m]$ with n_j annotated IID samples $(x_i, y_i) \overset{d}{\sim} p_{s_j}(x, y),\ i \in [n_j]$. The support set contains k-shot samples per class in the target domain, namely, $(x_i, y_i) \overset{d}{\sim} p_t(x, y),\ i \in [k \times |\mathcal{Y}^t|]$. For testing, we have a test set with n_t samples in the target domain. Our goal is to learn a predictor $f(\cdot)$ so the expected loss $\bar{L}_t = \mathbb{E}_{(x,y)\sim p_t} \mathbb{1}_{\{f(x) \neq y\}}$ is small. The k-shot samples in the support set are insufficient to learn a model for the new target space from scratch so the problem calls for techniques that can generalize from source datasets.

Zero-Shot Learning (ZSL): Again $p_s(x, y) \neq p_t(x, y)$ as in FSL. But in contrast to FSL we do not see annotated examples from target domain to help make a prediction.

For training, a sub-collection, $\mathcal{Y}^s \subset \mathcal{Y}$ of so called seen classes are only available and no other data from unseen class, i.e., no input data associated with $\mathcal{Y} \setminus \mathcal{Y}^s$ are available. To help train predictors, semantic vectors, $\sigma_y \in \Sigma$ for $y \in \mathcal{Y}^s$ are provided and the semantic vectors and labels are in one-to-one correspondence. The source distribution is characterized as $p_s(x, \sigma) \propto p(x, \sigma) \mathbb{1}_{\{\sigma_y : y \in \mathcal{Y}^s\}}$ and we obtain n_s IID instances $(x_i, \sigma_{y_i}) \sim p_s,\ i \in [n_s]$ for training. At test time we have full access to the semantic set Σ. In ZSL given an input instance $x \in \mathcal{X}$ from target unseen set namely the associated label $y \in \mathcal{Y}^t = \mathcal{Y} \setminus \mathcal{Y}^s$, our goal is to train a predictor, $f(x)$ that minimizes expected loss: $\bar{L}_t = \mathbb{E}_{(x,\sigma)\sim p_t} \mathbb{1}_{\{f(x) \neq \sigma\}}$, where $p_t \propto p(x, \sigma) \mathbb{1}_{\{\sigma_y : y \notin \mathcal{Y}^s\}}$.

Generalized Zero-Shot Learning (GZSL): While the training setup is identical to ZSL, at test-time, the input instances can be drawn from both seen and unseen object classes. Our goal is to minimize $\bar{L}_t = \mathbb{E}_{(x,\sigma)\sim p} \mathbb{1}_{\{f(x) \neq \sigma\}}$.

3.2. Overview of Proposed Approach

The overall structure of our model is illustrated by Fig.1. The proposed model consists of a cascade of functions, including a part-feature extractor, a part-probability encoder, and a task specific predictor designed for different applications, e.g. GZSL, FSL and DA.

Specifically, let $\mathcal{D}_{tr}$ denote the ordered pair of available input-output training instances, and $\mathcal{X}_{tr}$ the corresponding input training instances. For each input instance x, the part-feature extractor outputs M attention regions and associated features, $z(x) = [z_m(x)]_{m\in[M]}, z_m(x) \in \mathbb{R}^C$, where the attention regions focus on different foreground object parts and have negligible overlap in the image space. For each part, the part-probability encoder aims to discover K proto-typical atoms among the part-features in $\mathcal{X}_{tr}$, and project each part feature vector z_m on to such a dictionary of atoms D_m, resulting in a probability vector $\pi_m(x) \in \mathbb{R}^K, K \ll C$. The collection of part probability vectors $\pi(x) = [\pi_m(x)]_{m\in[M]}$ is then input into the task

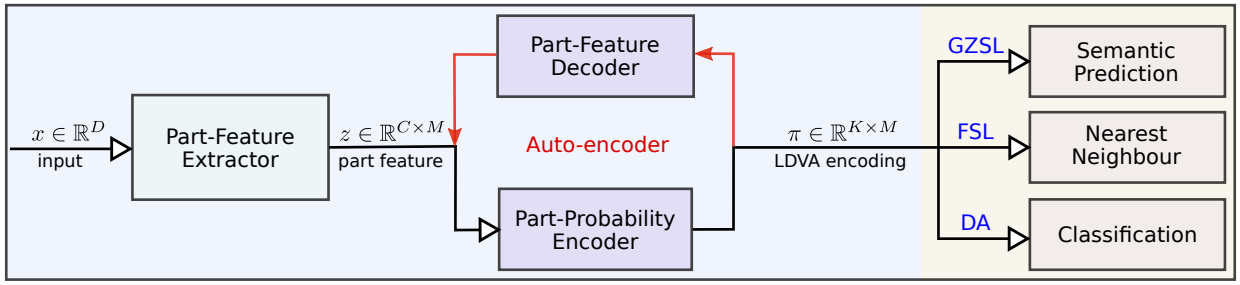

Figure 1. The proposed network architecture. For an input image x, the part feature extractor decompose x into M parts and extracts associated features z_m, the part-probability encoder then encodes each part feature as a low-dimensional encoding π_m by projecting the features onto a dictionary of primitive proto-typical part-types automatically discovered by our model. π_m is then used as inputs to train task-specific predictor models for GZSL, FSL and DA.

specific predictor $V(\pi(x))$, which outputs a class label.

The system is then trained to enforce three objectives: (1) the part-feature extractor should output diverse and discriminative attention regions that focus on common object parts prevalent on most instances in $\mathcal{X}_{tr}$; (2) the primitive prototypical atoms should be representative to reconstruct the original part-features; and finally, (3) the predictor should be customized and optimized for each specific task.

Prototypical Part Mixture Representation. We build intuition into how our proposed scheme leads to good generalization on the proposed problems. As such, each atom in the dictionary can be viewed as a prototypical part-type. Specifically, we assume part-features z_m in PTE can be represented by a Gaussian mixture of part-types. In other words, $z_m \sim \sum_k \pi_{k,m} \mathcal{N}(D_{k,m}, \gamma^2 I)$, where $\pi_{k,m}$ represents the probability part m belongs to component k of the Gaussian component $D_{k,m}$ as shown in Fig. 2.

Conditional Independence. From a probabilistic perspective we are placing a Markov chain structure on the relationship between input and output random variables (where, following convention, upper-case letters denote random variables): $X \longleftrightarrow \pi(X) \longleftrightarrow Y$. Thus $p(y \mid \pi(x), x) = p(y \mid \pi(x))$ and so $\pi(x)$ serves as a sufficient statistic and the only uncertainty that remains is to identify the prediction map from $\pi(x)$ to the labels $y \in \mathcal{Y}$ at test-time.

Discussion: How is the mixture representation effective? Suppose $p_s(x \mid y) \neq p_t(x \mid y)$, we could attempt to learn a mapping, $T(x)$ on the high-dimensional feature space so that for $x \sim p_t$ we have $T(x) \sim p_s(x)$, which we view as somewhat difficult. On the other hand, our proposed method learns and freezes LDVA encoding representing a composition of parts and offers benefits.

A. *Low-dimensionality.* The backbone network producing LDVA encoding is frozen at test-time. So learning a predictor on $\pi(x)$ requires relatively fewer examples.
B. *Compositional Uniqueness.* Attention regions are sufficiently representative of important aspects of objects in terms of discriminability of objects (Zheng et al., 2017). When the associated dictionary for each attention region are sufficiently descriptive, our visual encoding in terms of mixture composition uniquely describes different classes.
C. *Inter and Intra-Class Variances.* Intra-class variance arises from variance in visual appearance of a part-type within the same class and manifests in terms of the strength of the presence of the part-type in the input instance. On the other hand inter-class variance arises from the absence of parts or part-types, which results in smaller similarity in the visual encoding (see *Figure 3*).

Benefits of LDVA Encoding. In particular, under UDA we get $\pi(x) \approx \pi(x')$ for $x \sim p(\cdot|y)$, $x' \sim p(\cdot|y)$ requiring no further alignment. In FSL, we see new objects but as a consequence of (**B.**) these new objects are unique in terms of composition and furthermore as a result of (**C.**) are well separated. In ZSL we are given semantic vectors. Nevertheless, due to (**B.**) our representation closely mirror semantic vectors. Indeed, for many datasets, human-labeled semantic components are based on presence of visual parts in the class, and thus well-matched to LDVA encoding.

3.3. Model and Loss Parameterization

Part-Feature Extractor: Inspired by (Zheng et al., 2017), we use a multi-attention convolutional neural network (MA-CNN) to map input images into a finite set of part feature vectors, $z_m \in \mathbb{R}^C$. Specifically, it contains a global feature extractor E and a channel grouping model G, where $E(x) \in \mathbb{R}^{W \times H \times C}$ is a global feature map, and $G(E(x)) \in \mathbb{R}^{M \times C}$ is a channel grouping weight matrix. We then calculate an attention map $A_m(x) \in \mathbb{R}^{W \times H}$ for the m-th part:

$$A_m(x) = \text{sigmoid}\big(\sum_c G_{m,c}(x) \times E_c(x)\big) \qquad (1)$$

The part feature $z_m \in \mathbb{R}^C$ is then calculated as:

$$z_{m,c} = \sum_{w,h} [A_m(x) \odot E_c(x)]_{(w,h)}, \quad \forall c \in [C] \qquad (2)$$

where $\odot$ is the element-wise multiplication. We parameterized $E(\cdot)$ by the ResNet-34 backbone (to $conv5\_x$), and $G(\cdot)$ by a fully-connected layer.

To encourage a part-based representation z_m to be learned,

we follow (Zheng et al., 2017). Since z_m can be decomposed into $A_m(x) \odot E(x)$, we want to force the learned attention maps A_m to be both compact within the same part, and divergent among different parts. We define ℓ_{part} to be:

$$\ell_{part}(x) = \sum_m (L_{dis}(A_m(x)) + \lambda L_{div}(A_m(x))) \quad (3)$$

where the compact loss $L_{dis}(A_m)$ and divergent loss $L_{div}(A_m)$ are defined as (x is dropped for simplicity):

$$L_{dis}(A_m) = \sum_{w,h} A_m^{w,h}[\|w - w^*\|^2 + \|h - h^*\|^2] \quad (4)$$

$$L_{div}(A_m) = \sum_{w,h} A_m^{w,h}[\max_{n,n\neq m} A_n^{w,h} - \zeta] \quad (5)$$

where $A_m^{w,h}$ is the amplitude of A_m at coordinate (w, h), and (w^*, h^*) is the coordinate of the peak value of A_m, ζ is a small margin to ensure the training robustness.

Part-Probability Encoder: Our Gaussian assumption leads us to an auto-encoder implementation to map the high-dimensional part-feature $z_m(x) \in \mathbb{R}^C$ into the low-dimensional probability $\pi_m(x) \in \mathbb{R}^K$. Specifically, for part m, given the part features $z_m(x) \in \mathbb{R}^C$, we define a projection matrix $P_m \in \mathbb{R}^{K\times C}$, such that:

$$\pi_m(x) = P_m z_m(x) \quad (6)$$

Guassian Mixture Condition: Our Gaussian mixture assumption (Sec.3.2) implies the following condition should hold:

$$z_m(x) \approx D_m^\top \pi_m(x), \quad (7)$$

where $D_m \in \mathbb{R}^{K\times C}$ is a 'fat' matrix ($K \ll C$) of Gaussian components for part m, i.e. $D_m = [D_{k,m}]_{k\in[K]}$.

Viewing P_m and D_m as model parameters, our training objective can be naturally written in the form of an auto-encoder, where P_m is the encoder and D_m is the decoder:

$$\ell_{prob}(x) = \textstyle\sum_m \left(\|z_m(x) - D_m^\top P_m z_m(x)\|^2 + \lambda\|P_m\|^2 + \lambda\|D_m\|^2\right). \quad (8)$$

Task Specific Predictors: The part-probability π serves as an input to a task specific predictor $V(\pi)$.

Generalized Zero-Shot Learning: For GZSL, $V(\pi)$ is a semantic prediction model parameterized by a neural network to project π into the semantic space Σ. Given an input image x and its semantic attribute σ_y, the loss for training the GZSL predictor, with η as margin parameter is modeled as:

$$\ell_{GZSL}(x, y) = \textstyle\sum_{y'\in\mathcal{Y}} [\eta \mathbb{1}[y' = y] + \sigma_y^\top V(\pi(x)) - \sigma_y^\top V(\pi(x))]_+ \quad (9)$$

Few-Shot Learning: For FSL, we have different implementations for the predictor in the source domain and the target domain. For an input-output pair (x, y) in the source domain training set, $V(\pi)$ is a classification model parameterized by a neural network to project π into the class label space $\mathcal{Y}^s$. The training loss is simply a cross-entropy loss:

$$\ell_{FSL}(x, y) = \text{CE}(V(\pi(x)), o(y)), \quad (10)$$

where $\text{CE}(\cdot,\cdot)$ is the cross-entropy, $o(\cdot)$ is the one-hot encoding function. After training, we calculate the average $\bar{\pi}_y$ representation for K-shot samples in the target domain support set, and build a nearest neighbour classifier for testing, i.e. $V(\pi(x)) = \arg\min_{y\in\mathcal{Y}^t} |\pi(x) - \bar{\pi}_y|^2$.

Domain Adaptation: For DA, $V(\pi)$ is a classification model parameterized by a neural network to project π into the class label space $\mathcal{Y}$. In DA we have training samples from both source domain $\mathcal{D}_s$ and target domain $\mathcal{D}_t$, where D_t has no class label available during training. Inspired by (Chadha & Andreopoulos, 2018; Saito et al., 2017), we estimate psuedo-labels for the target domain samples with the current classification model $V(\pi)$ and further optimizes the following loss:

$$\begin{aligned}\ell_{DA}(x, y) =& \mathbb{1}[x \in \mathcal{D}_s]\text{CE}(V(\pi(x)), o(y)) \\ &+ \mathbb{1}[x \in \mathcal{D}_t]\text{CE}(V(\pi(x)), o(\hat{y})),\end{aligned} \quad (11)$$

where $\hat{y} = \arg\max_y V(\pi(x))_y$, and $V(\pi(x))_y$ is the y-th element in the $V(\pi(x))$ vector.

By pseudo-labelling target samples, we aim to align the class level source-target distributions, i.e. aligning $p_s(x|y)$ and $p_t(x|y)$, and meanwhile minimize the entropy of the prediction distributions, such that a discriminative π representation that convey confident decision rules can be learnt.

End-to-End Training: We train our system discriminatively by employing three loss functions. In particular, suppose the part-feature extractor is parameterized by Θ, the part-probability encoder by $([D_m, P_m])$, the predictor by α, the overall training objective is:

$$\begin{aligned}\min_{\theta,\alpha,[D_m,P_m]} \sum_{(x,y)\in\mathcal{D}_{tr}} &\ell_{part}(x;\Theta) + \ell_{prob}(x;[D_m, P_m],\Theta) \\ &+ \ell_{task}(x, y; \alpha, [D_m, P_m], \Theta)\end{aligned} \quad (12)$$

where $\ell_{task} \in \{\ell_{GZSL}, \ell_{FSL}, \ell_{DA}\}$.

3.4. Implementation Details

We set the number of parts M to 4 and in each part the number of prototypes K is set to 16. ζ in Eq.(5) is empirically set to 0.02. For FSL, we set the input size to be [224 224], and λ in Eq.(3) is 2; for GZSL, our model takes input image size as [448 $\times$ 448] and λ is set to 5; For DA, the input image size is [224 $\times$ 224] and λ is set to 2. The task-specific predictor $V(\cdot)$ for both GZSL and DA is implemented by a two FC-layer neural network with ReLU activation, the number of neurons in the hidden layer is set to 32.

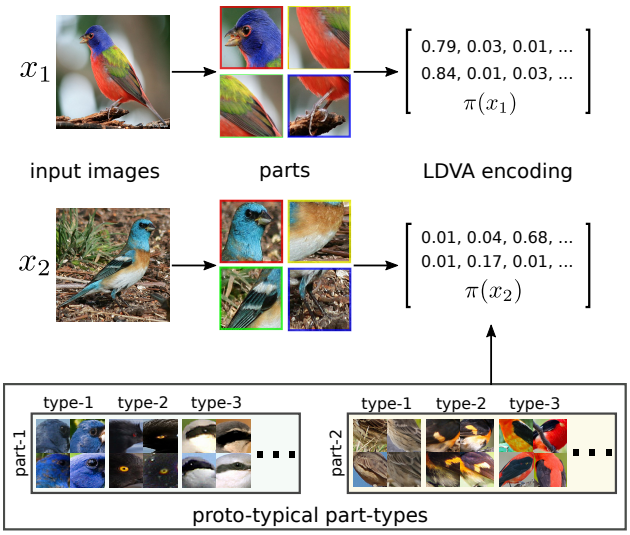

Figure 2. LDVA is generated so that the mixture of the proto-typical part types represents the corresponding part. The objects have similar representation if they have similar visual parts. The resulting LDVA encoding also has a smaller gap to the semantic attributes in the GZSL setting, e.g. beak-color, wing-color, etc, compared to the original high-dimensional features.

Our model takes an alternative optimization approach to minimize the overall loss. In each epoch, we update the weights in two steps. In step.A, only the weights of channel grouping $G(\cdot)$ is updated by minimizing ℓ_{part}. In step.B, we freeze the weights of $G(\cdot)$ and update all the other modules. Adam optimizer is used in each step.

4. Experiments

4.1. Few-Shot Learning

Datasets. We first evaluate the few shot learning performance of the proposed model on two benchmark datasets: Omniglot(Lake et al., 2015) and *mini*ImageNet(Vinyals et al., 2016). Omniglot consists of 1623 characters from 50 alphabets. Each character (class) contains 20 handwritten images from people. *mini*ImageNet is a subset of ImageNet(Russakovsky et al., 2014) which contains 60,000 images from 100 categories.

Setup. We follow the same protocol in (Sung et al., 2018). For Omniglot, the dataset is augmented with new classes through 90°, 180° and 270° rotations of existing characters. 1200 original classes plus rotations are selected as training set and the remaining 423 classes with rotations are test set. For *mini*ImageNet, the dataset is split into 64 training, 16 validation and 20 testing classes. The model will only be trained on training set and the validation set is for examining the training performance.

We evaluate the 5-way accuracy on *mini*ImageNet and 5-way plus 20-way accuracy on Omniglot. 1-shot and 5-shot learning performance is evaluated in each setting. For *m-way k-shot* learning, in each test episode, m classes will be randomly selected from the test set, then k samples will be drawn from these classes as support examples, and 15 examples will be drawn from the rest images to construct the test set. We run 1000 and 600 test episodes on Omniglot and *mini*ImageNet, respectvely, to compute the average classification accuracy.

Methods	Omniglot				*mini*ImageNet	
	5-way Acc.		20-way Acc.		5-way Acc.	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML	98.7	**99.9**	95.8	98.9	48.7	63.1
Prototypical Nets	98.8	99.7	96.0	98.9	49.4	68.2
Relation Nets	**99.6**	99.8	**97.6**	99.1	50.4	65.3
TADAM	-	-	-	-	58.5	76.7
LEO	-	-	-	-	61.7	77.6
EA-FSL	-	-	-	-	**62.6**	78.4
Ours	98.9	99.8	96.5	**99.3**	61.7	**78.7**

Table 1. FSL classification results on Omniglot and *mini*ImageNet.

Training Details. Our model is trained for 80 and 30 epochs on Omniglot and *mini*ImageNet, repectively. The learning rate for step.A is set to 1e-6, and the learning rate of step.B is 1e-4 for Omniglot and 1e-5 for *mini*ImageNet. On *mini*ImageNet, the weights for the feature extractor $E(\cdot)$ is pretrained on the training split.

Competing Models. We list here the state-of-the-art methods we compare to: Prototypical Nets(Snell et al., 2017), MAML(Finn et al., 2017), Relation Nets(Sung et al., 2018), TADAM(Oreshkin et al., 2018), LEO(Rusu et al., 2018), and EA-FSL(Ye et al., 2018).

Results. Few shot learning results are shown in *Table 1*. On both datasets, our model reaches the same level accuracy as other state-of-the-art methods. Specifically, on *mini*ImageNet, our model obtain 78.7% for 5-shot learning scanerio, which supass the second best model with an absolutely margin 0.3%. On omniglot, the accuracy for 20-way 5-shot learning is improved to 99.3%.

Compared with other methods which process the high-dimensional visual features or utilize meta-learning strategy, our model leverages the LDVA representations to reduce the inter-class variance for novel categories. That is, because of the unique composition for each class, the distance between examples in the same class is smaller than the high-dim features. This results in the good performance in LDVA even only a simple nearest neighbor classifier is applied. In addition, since a universal part prototypes is learned from the seen classes, our model does not require any meta-training or fine-tune on the unseen categories, while meat-learning based methods need to dynamically adapt their model based on the feedback from new tasks.

4.2. Generalized Zero-Shot Learning

Datasets. The performance of our model for GZSL is evaluated on three commonly used benchmark datasets: *Caltech-*

Methods	CUB			AWA2			aPY		
	ts	tr	H	ts	tr	H	ts	tr	H
SSE(Zhang & Saligrama, 2015)	8.5	46.9	14.4	8.1	82.5	14.8	0.2	**78.9**	0.4
ALE(Akata et al., 2016)	23.7	62.8	34.4	14.0	81.8	23.9	4.6	73.7	8.7
SYNC(Changpinyo et al., 2016)	11.5	70.9	19.8	10.0	90.5	18.0	7.4	66.3	13.3
DEVISE(Frome et al., 2013)	23.8	53.0	32.8	17.1	74.7	27.8	4.9	76.9	9.2
PSRZSL(Annadani & Biswas, 2018)	24.6	54.3	33.9	20.7	73.8	32.3	13.5	51.4	21.4
SP-AEN(Chen et al., 2018)	34.7	70.6	46.6	23.3	**90.9**	37.1	13.7	63.4	22.6
GDAN(Huang et al., 2018)	39.3	66.7	49.5	32.1	67.5	43.5	**30.4**	75.0	**43.4**
CADA-VAE(Schönfeld et al., 2018)	51.6	53.5	52.4	**55.8**	75.0	**63.9**	-	-	-
SE-GZSL(Kumar Verma et al., 2018)	41.5	53.3	46.7	**58.3**	68.1	62.8	-	-	-
LSD(Dong et al., 2018)	**53.1**	59.4	**56.1**	-	-	-	22.4	**81.3**	35.1
Ours	33.4	**87.5**	48.4	41.6	**91.3**	57.2	24.5	72.0	36.6
Ours + CS	**59.2**	**74.6**	**66.0**	54.6	87.7	**67.3**	**41.1**	68.0	**51.2**

Table 2. GZSL results on CUB, AWA2 and aPY. ts = test classes (unseen classes), tr = train classes (seen classes), H = harmonical mean. The accuracy is class-average Top-1 in %. The highest accuracy is in red color and the second is in blue (better viewed in color).

UCSD Birds-200-2011 (CUB) (Wah et al., 2011), *Animals with Attributes 2* (AWA2) (Xian et al., 2018a) and Attribute Pascal and Yahoo (aPY) (Farhadi et al., 2009). CUB is a fine-grained dataset consisting of 11,788 images from 200 different types of birds. 312-dim semantic attributes are annotated for each category. AWA2 is a coarse-grained dataset which has 37,222 images from 50 different animals and 85-dim class-level semantic attributes. aPY contains 20 Pascal classes and 12 Yahoo classes. It has 15,339 images in total and 64-dim semantic attributes are provided.

Setup. Recent works (Xian et al., 2018a) have shown that the conventional ZSL setting is overly optimistic because it leverages absence of seen classes at test-time and there is an emerging consensus that methods should focus on the generalized ZSL setting. We thus evaluated under the GZSL setting. Following the protocol in (Xian et al., 2018a), we evaluate the average-class Top-1 accuracy on unseen classes (ts), seen classes (tr) and the harmonic mean (H) of ts and tr.

It has been observed the scores for seen classes are often greater than unseen in GZSL methods (Chao et al., 2016), which results in poor performance. Calibrated Stacking(CS) is proposed in (Chao et al., 2016) to balance the performance between seen and unseen classes by calibrating the scores of seen classes. As tabulated in *Table 2*, in addition to our original model, we also apply CS into our model to alleviate this imbalance, denoted as (Ours+CS). The parameters for CS is chosen via cross validation.

Training Details. Our models are trained for 120, 100 and 110 epochs on CUB, AWA2 and aPY, respectively. The learning rate for step.A and step.B is set to 1e-6 and 1e-5.

Competing Models. We compare against state-of-the-art approaches: ALE(Akata et al., 2016), DEVISE(Frome et al., 2013), SSE(Zhang & Saligrama, 2015), SYNC(Changpinyo et al., 2016), PSRZSL(Annadani & Biswas, 2018), SP-AEN(Chen et al., 2018), GDAN(Huang et al., 2018), CADA-VAE(Schönfeld et al., 2018), SE-GZSL(Kumar Verma et al., 2018), and LSD(Dong et al., 2018). Recently there are some methods in trasductive ZSL setting (Zhao et al., 2018; Zhang et al., 2018a) where the data for unseen classes are accessible during training. Although it is interesting to leverage the unseen data and boost the performance for gZSL(Zhang et al., 2018a), we focus on learning a universal representation (LDVA) for both seen and unseen classes from limited data. Therefore we only compare with inductive methods listed above.

Results. Results for GZSL are in *Table 2*. Without the calibrated stacking, our model (ours) reaches 48.4% on CUB, 57.2% on AWA2 and 36.6% on aPY for the harmonic mean (H). After the scores are calibrated, our model (ours+CS) obtains 66.0%, 67.3% and 51.2% for the harmonic mean, respectively, which outperforms all other competing models. Specifically, on CUB, Ours+CS surppasses the 2-nd best result (LSD) by a margin of 6.1% on ts, 15.2% on tr, and 9.9% on H. On AWA2, our models increase the accuracy on H from 63.9% to 67.3%. On aPY, our models improves the accuracy for unseen classes from 30.4% to 41.1%, resulting in a 7.8% increase on harmonic mean.

The success of our model can be attributed primarily to the proposed LDVA representation, which resembles the components of the semantic attributes. For example, in *Figure 2*, we visualize the part attentions discovered by our model and several semantic attributes for the class *'Painted_Bunting'* in CUB dataset. Our model learns the part areas around "head", "wing", "body", and "feet", based on which are most semantic attributes annotated (e.g. crow color: blue, wing color: green, etc.). Via the prototype encoding, our visual attributes mirror the representation of semantic vectors, thus mitigating the large gap between the semantic attributes and high-dimensional visual features.

4.3. Domain Adaptation

Datasets. We evaluate our proposed model in unsupervised domain adaptation task between three digits datasets: MNIST(LeCun et al., 1998), USPS and SVHN(Netzer et al., 2011). Each dataset contains 10 classes of digit numbers (0-9). MNIST and USPS are handwritten digits while SVHN is obtained from house number in google street view images.

Methods	M → U	U → M	S → M
CoGAN	91.2	89.1	-
ADDA	89.4	90.1	76.0
UNIT	96.0	93.6	90.5
CyCADA	95.6	96.5	90.4
MSTN	92.9	-	91.7
Ours (source π)	94.8	96.1	82.4
Ours (joint π)	**98.8**	**96.8**	**95.2**

Table 3. DA classification results. M = MNIST, U = USPS, S = SVHN. The highest accuracy is in **bold**

Setup. We follow the same protocol in (Tzeng et al., 2017), where the adaptation in three directions are validated: MNIST→USPS, USPS→MNIST, and SVHN→MNIST. In the experiments, two variants of our model are evaluated: (1) **Ours(source π)**: During training, the model is purely learned from source data. In this case, ℓ_{DA} reduces to a standard cross-entropy loss on the source domain. In test time, LDVA encoding for target data is based on the source visual encoder P_m^s. This model does not utilize any information from the unlabeled target data in the training. (2) **Ours(joint π)**: This model learns the visual encoder from the joint dataset $\mathcal{D}_s \cup \mathcal{D}_t$ as described by Eq.(11).

Training Details. Ours (source π) is trained on the source domain dataset, as described above. The learning rate for step.A and step.B is 1e-6 and 1e-5. The training epochs are set to be 40, 20, and 40 on MNIST, USPS, and SVHN, respectively. For joint π, we first initialize with weights from trained on our source-π model. Next, the model is trained on the joint dataset $\mathcal{D}_s \cup \mathcal{D}_t$ for 10 epochs. The learning rate for step.B is modified to 1e-6.

Competing Models. We compare against several state-of-the-art UDA methods: ADDA (Tzeng et al., 2017), CoGAN (Liu & Tuzel, 2016), UNIT (Liu et al., 2017), CyCADA (Hoffman et al., 2017), MSTN (Xie et al., 2018), and Self-ensembling(French et al., 2017).

Results. The results for DA are shown in *Table 3*. Specifically, Ours (source π) and Ours (joint π) reaches 94.8%, 98.8% on M→U, 96.1%, 96.8% on U→M, and 82.4%, 95.2% on S→M. Observe that our method with jointly learned π outperforms all other competing methods, with a margin of 3.5% on S→M, 2.8% on M→U, and 0.4% on U→S compared to the second-best competitors.

The results demonstrate the benefits of proposed LDVA representation. Specifically, in the same domain, the LDVA representations for different classes are large enough to learn a good classifier. Meanwhile, the representations of the same class from different domains are much more similar than the high-dimensional features, resulting in a similar distribution for π_s and π_t. The classifier on source domain are thus able to be applied on target domain. We also illustrate this effect in Fig.3(a-b). As we can see the part probability vector of digit '2' in MNIST is very similar to SVHN, while quite different against the digit '5' in MNIST.

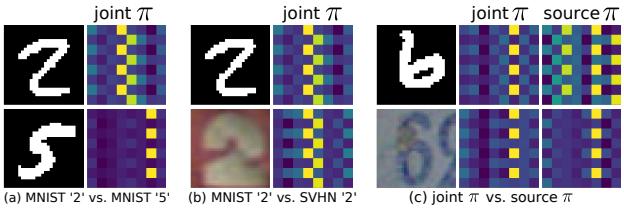

Figure 3. Proposed LDVA encoding π on digit datasets. The 64-dimensional π vector is reshaped to a 8×8 matrix for better visualization. For all three examples (a-c), π is trained for SVHN→MNIST experiment.

Source vs. Joint π. Our model with jointly learned π outperforms purely source π with 4.4%, 0.7% and 12.8% absolute improvement on the three adaptation directions. This comparison shows that the cross-entropy loss using pseudo-label on the target domain in Eq.(11) helps the model learn a more universal prototypes, and hence reduces the distance between the representations of the same class. The model will benefit more when the domain shift is severe. As shown in Fig.3(c), the part probability vector of digit '6' in MNIST is more similar to the one in SVHN in the joint π space. This also results in the largest performance gap on S→M, since SVHN is obtained from street view while MNIST and USPS are both handwritten digits.

Tolerance to Visual Distortions. Note that all of the competing methods are trained jointly on both source and target domains and so, comparing against our source-π method is an unfair comparison. Still, what we see here from the first two experiments is that access to unlabeled target data is somewhat unnecessary if we adopt LDVA encoding. This points to the fact that mixture compositions are tolerant to visual distortions, which can be an issue for methods relying on transferring information in high-dimensions. On the other hand for the last experiment, the variance is significant and unannotated target data is useful.

5. Conclusion

We proposed a novel method for computer vision problems, where new tasks and environments arise. In these cases, due to limited supervision on the target, training "models from scratch," is impossible and methods that adapt existing models, trained on the presented training environment, to the new scenario are required. We propose a novel low-dimension visual attribute (LDVA) encoding method that represents the mixture composition of prototypical parts of any instance. The LDVA encodings are low-dimensional, are capable of uniquely representing new objects and are tolerant to visual distortions. We train an end-to-end model for a variety of tasks including domain adaptation, few shot learning and zero-shot learning. Our method outperforms state-of-art methods even though those methods are customized to the specific problem contexts (ZSL, FSL, DA).

References

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2016.

Annadani, Y. and Biswas, S. Preserving semantic relations for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Antol, S., Zitnick, C. L., and Parikh, D. Zero-shot learning via visual abstraction. In *ECCV*, pp. 401–416. Springer, 2014.

Antoniou, A., Storkey, A., and Edwards, H. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Bhatia, K., Jain, H., Kar, P., Varma, M., and Jain, P. Sparse local embeddings for extreme multi-label classification. In *NIPS*, 2015.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 7, 2017.

Chadha, A. and Andreopoulos, Y. Improving adversarial discriminative domain adaptation. *arXiv preprint arXiv:1809.03625*, 2018.

Changpinyo, S., Chao, W.-L., Gong, B., and Sha, F. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5327–5336, 2016.

Chao, W.-L., Changpinyo, S., Gong, B., and Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016.

Chen, L., Zhang, H., Xiao, J., Liu, W., and Chang, S.-F. Zero-shot visual recognition using semantics-preserving adversarial embedding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2018.

Chen, Y., Wang, J., Bai, Y., Castanon, G., and Saligrama, V. Probabilistic semantic retrieval for surveillance videos with activity graphs. *IEEE Transactions on Multimedia*, 21(3):704716, Mar 2019.

Ding, Z., Li, S., Shao, M., and Fu, Y. Graph adaptive knowledge transfer for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 36–52. Springer, 2018.

Dong, H., Fu, Y., Sigal, L., Hwang, S. J., Jiang, Y.-G., and Xue, X. Learning to separate domains in generalized zero-shot and open set learning: a probabilistic perspective. *arXiv preprint arXiv:1810.07368*, 2018.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1778–1785. IEEE, 2009.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

French, G., Mackiewicz, M., and Fisher, M. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., and Li, W. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

Huang, H., Wang, C., Yu, P. S., and Wang, C.-D. Generative dual adversarial network for generalized zero-shot learning. *arXiv preprint arXiv:1811.04857*, 2018.

Jiang, H., Wang, R., Shan, S., and Chen, X. Learning class prototypes via structure alignment for zero-shot recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Kang, G., Zheng, L., Yan, Y., and Yang, Y. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. *arXiv preprint arXiv:1801.10068*, 2018.

Kodirov, E., Xiang, T., and Gong, S. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.

Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., and Wornell, G. Co-regularized alignment for unsupervised domain adaptation. In *Advances in*

Neural Information Processing Systems, pp. 9367–9378, 2018.

Kumar Verma, V., Arora, G., Mishra, A., and Rai, P. Generalized zero-shot learning via synthesized examples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, C.-W., Fang, W., Yeh, C.-K., and Frank Wang, Y.-C. Multi-label zero-shot learning with structured knowledge graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Li, Y., Zhang, J., Zhang, J., and Huang, K. Discriminative learning of latent features for zero-shot recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Lin, X., Wang, H., Li, Z., Zhang, Y., Yuille, A., and Lee, T. S. Transfer of view-manifold learning to similarity perception of novel objects. *arXiv preprint arXiv:1704.00033*, 2017.

Liu, M.-Y. and Tuzel, O. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pp. 469–477, 2016.

Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pp. 700–708, 2017.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1647–1657, 2018.

Mehrotra, A. and Dukkipati, A. Generative adversarial residual pairwise networks for one shot learning. *arXiv preprint arXiv:1703.08033*, 2017.

Munkhdalai, T. and Yu, H. Meta networks. *arXiv preprint arXiv:1703.00837*, 2017.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Oreshkin, B., López, P. R., and Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 721–731, 2018.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge, 2014.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

Saito, K., Ushiku, Y., and Harada, T. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017.

Schönfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., and Akata, Z. Generalized zero-and few-shot learning via aligned variational autoencoders. *arXiv preprint arXiv:1812.01784*, 2018.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2242–2251. IEEE, 2017.

Shu, R., Bui, H. H., Narui, H., and Ermon, S. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

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

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 4, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

Wang, J., Feng, W., Chen, Y., Yu, H., Huang, M., and Yu, P. S. Visual domain adaptation with manifold embedded distribution alignment. In *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 402–410. ACM, 2018a.

Wang, X., Ye, Y., and Gupta, A. Zero-shot recognition via semantic embeddings and knowledge graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

Wang, Y.-X., Girshick, R., Hebert, M., and Hariharan, B. Low-shot learning from imaginary data. *arXiv preprint arXiv:1801.05401*, 8, 2018c.

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

Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. Feature generating networks for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018b.

Xie, S., Zheng, Z., Chen, L., and Chen, C. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 5419–5428, 2018.

Ye, H.-J., Hu, H., Zhan, D.-C., and Sha, F. Learning embedding adaptation for few-shot learning. *arXiv preprint arXiv:1812.03664*, 2018.

Zhang, L., Wang, P., Liu, L., Shen, C., Wei, W., Zhang, Y., and Hengel, A. V. D. Towards effective deep embedding for zero-shot learning, 2018a.

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

Zhang, Z. and Saligrama, V. Zero-shot recognition via structured prediction. In *European conference on computer vision*, pp. 533–548. Springer, 2016.

Zhang, Z., Liu, Y., Chen, X., Zhu, Y., Cheng, M.-M., Saligrama, V., and Torr, P. H. Sequential optimization for efficient high-quality object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1209–1223, 2018b.

Zhao, A., Ding, M., Guan, J., Lu, Z., Xiang, T., and Wen, J.-R. Domain-invariant projection learning for zero-shot recognition. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1019–1030. Curran Associates, Inc., 2018.

Zheng, H., Fu, J., Mei, T., and Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*, volume 6, 2017.

Zhu, P., Wang, H., and Saligrama, V. Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 11, 2019a.

Zhu, P., Wang, H., and Saligrama, V. Generalized zero-shot recognition based on visually semantic embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., and Elgammal, A. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Acknowledgement

The authors would like to thank the Area Chair and the reviewers for their constructive comments. This work was supported by the Office of Naval Research Grant N0014-18-1-2257, NGA-NURI HM1582-09-1-0037 and the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Grant 2013-ST-061-ED0001.