Structured Latent Embeddings for Recognizing Unseen Classes in Unseen Domains

Shivam Chandhok[1]
shivam.chandhok@mbzuai.ac.ae
Sanath Narayan[2]
sanath.narayan@inceptioniai.org
Hisham Cholakkal[1]
hisham.cholakkal@mbzuai.ac.ae
Rao Muhammad Anwer[13]
rao.anwer@mbzuai.ac.ae
Vineeth N Balasubramanian[4]
vineethnb@iith.ac.in
Fahad Shahbaz Khan[15]
fahad.khan@mbzuai.ac.ae
Ling Shao[2]
ling.shao@ieee.org

[1] Mohamed bin Zayed University of AI, UAE
[2] Inception Institute of Artificial Intelligence, UAE
[3] Aalto University School of Science, Espoo, Finland
[4] Indian Institute of Technology, Hyderabad, India
[5] Linköping University, Sweden

Abstract

Zero-shot learning and domain generalization strive to overcome the scarcity of task-specific annotated data by individually addressing the issues of semantic and domain shifts, respectively. However, real-world applications often are unconstrained and require handling unseen classes in unseen domains, a setting called zero-shot domain generalization, which presents the issues of domain and semantic shifts simultaneously. Here, we propose a novel approach that learns domain-agnostic structured latent embeddings by projecting images from different domains and their class-specific semantic representations to a common latent space. Our method jointly strives for the following objectives: (i) aligning the multimodal cues from visual and text-based semantic concepts; (ii) partitioning the common latent space according to the domain-agnostic class-level semantic concepts; and (iii) learning a domain invariance w.r.t the visual-semantic joint distribution for generalizing to unseen classes in unseen domains. Our experiments on challenging benchmarks such as DomainNet show the superiority of our approach over existing methods with significant gains on difficult domains like *quickdraw* and *sketch*.

1 Introduction

In various real-world computer vision problems, obtaining task-specific labeled data can be challenging due to several reasons: high annotation costs, dynamic addition of objects with

new semantic content or in new domains, limited instances of rare objects, as noted in [35]. Two popular paradigms addressing such issues are: (i) zero-shot learning (ZSL) that employs training data of related object categories from the same domain (*e.g.*, *sketches* of cats as training data for recognizing dogs in *sketches*); and (ii) domain generalization (DG) that uses training data of the same categories from related domains (*e.g.*, *photos* of dogs as training data for recognizing dogs in *sketches*). While ZSL tackles the semantic shift caused by different object categories during training and testing, DG handles the domain shift caused by different domains in training and testing data. However, real-world applications often require handling semantic and domain shifts simultaneously. Here, we investigate this challenging problem of zero-shot domain generalization (ZSLDG) [26, 28], which leverages training data of a related object category from a related domain for recognizing unseen classes in unseen domains (*e.g.*, *photos* of cats as training data for recognizing dogs in *sketches*).

In this work, we address ZSLDG by proposing a unified solution that jointly tackles domain and semantic shifts by relating the visual cues of a class to its domain-invariant semantic concepts. *E.g.*, semantic cues such as <*long neck*, *long legs*, *has spots*> of *giraffe* class are invariant across domains such as *real* images (photos), *sketch* or *clipart* (see Fig. 1). To this end, we align common information from visual and semantic spaces in a domain-agnostic latent space that is structured according to class-level semantic concepts. We then impose a domain invariance w.r.t the visual-semantic joint distribution. Since the semantic space is shared across all classes and agnostic to the visual domains, imposing such invariance aids in generalizing to unseen domains at test time (instead of overfitting to source domains), while improving the visual-semantic interaction for effective knowledge transfer across seen and unseen classes.

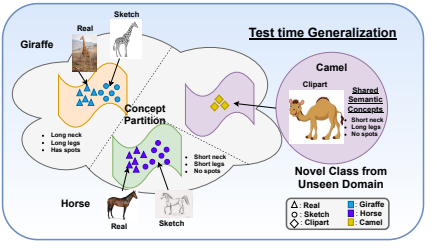

Figure 1: Our latent space is structured according to class-level semantic concepts and is domain-invariant w.r.t visual-semantic joint distribution. This enables our model to map unseen classes in unseen domains at test-time (*camel* from *clipart*), based on their semantic concepts, to appropriate subspaces within our latent space, thereby aiding generalization. Best viewed in zoom.

Contributions: We propose a ZSLDG approach comprising a visual encoder that projects multi-domain images from the visual space to a latent space, and a semantic encoder that learns to map text-based class-specific semantic representations to the same latent space. The key contributions are: **(i)** For aligning class-specific cues from visual and semantic latent embeddings, we introduce a multimodal alignment loss term; **(ii)** We partition the latent space w.r.t class-level semantic concepts across domains by minimizing intra-class variance across different seen domains; **(iii)** The focus of our design is the introduction of a joint invariance module that strives for domain invariance w.r.t the visual-semantic joint distribution, and thereby facilitates generalizing to unseen classes in unseen domains; and **(iv)** Experiments on the challenging DomainNet and DomainNet-LS benchmarks [36] demonstrate the efficacy of our approach over existing methods. Particularly, on the difficult *quickdraw* domain, our approach achieves a significant gain of 1.6% over the best existing method [26].

2 Related Work

Domain Generalization (DG): Existing methods tackle the problem of domain shift, which occurs when the training and testing data belong to different domains, in different ways.

Most previous approaches aim to learn domain-invariance by minimizing the discrepancy between multiple source domains [31, 49, 50] or by employing autoencoders and adversarial losses [13, 23]. A few works [5, 21, 22] introduce specific training policies or optimization procedures such as meta-learning and episodic training to enhance the generalizability of the model to unseen domains. Similarly, [39, 42] employ data augmentation strategies to improve the models robustness to data distribution shifts at test time. However, all these works tackle the DG problem alone, where the label spaces at both train and test time are identical.

Zero-shot Learning (ZSL): Traditional ZSL methods [1, 2, 11, 37, 40] learn to project the visual features onto a semantic embedding space via direct mapping or through a compatibility function. However, such direct mappings are likely to suffer from issues of seen class bias and hubness [9, 18]. In contrast, the works of [33, 41] leverage joint multi-modal learning of visual and textual feature embeddings for the task of ZSL. Recently, generative approaches tackle the problem of seen class bias by generating unseen visual features from respective class embeddings [6, 7, 10, 15, 24, 27, 30, 32, 38, 43, 45, 48]. However, all these methods address only ZSL, where the domain remains unchanged during training and testing.

Zero-shot Domain Generalization (ZSLDG): The recent works of [26, 28] investigate the challenging problem of ZSLDG. While [28] limits its real-world applicability by defining different domains as variations in rotation of the same objects, CuMix [26] defines domains as different ways of depicting an object, as in *sketch*, *painting*, *cartoon*, *etc*. and is closer to the real-world settings. CuMix tackles the issue of domain shift through data augmentation by mixing source domains, and handles semantic shifts by learning to project visual features to the semantic space. However, relying on source domain-mixing is likely to result in a model that could overfit to the source domains and their interpolations, thereby reducing generalizability to unseen domains [23]. Furthermore, directly mapping the visual space to the semantic space, as in [26], can lead to hubness issues (mapped points cluster as a hub due to low variance) [9, 18], thereby reducing class-discriminability. In contrast, our approach jointly handles the issues of domain and semantic shifts by learning a domain-agnostic latent space that is partitioned based on domain-invariant class-level semantic concepts, onto which the visual and semantic features are projected. Since enforcing domain-invariance w.r.t marginal distribution of images is likely lead to overfitting towards seen domains [23, 25], we tackle this by enforcing domain-invariance w.r.t the visual-semantic joint distribution. In addition, this enables better interaction between visual and semantic spaces in a new latent space, thereby improving the generalization to unseen classes in unseen domains.

3 Proposed Method

Problem Setting: The goal of zero-shot domain generalization (ZSLDG) is to recognize unseen classes in unseen domains. Let $Q^{Tr} = \{(\mathbf{x}, y, \mathbf{a}_y, d) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}^s, \mathbf{a}_y \in \mathcal{A}, d \in \mathcal{D}^s\}$ denote the training set, where $\mathbf{x}$ is a seen class image in the visual space ($\mathcal{X}$) with corresponding label y from a set of seen class labels $\mathcal{Y}^s$. Here, $\mathbf{a}_y$ denotes the class-specific semantic representation that encodes the inter-class relationships, while d is the domain label from a set of seen domains $\mathcal{D}^s$. Note that the semantic representations are typically obtained from unsupervised text-based WordNet models (*e.g.*, *word2vec* [29]) Similarly, $Q^{Ts} = \{(\mathbf{x}, y, \mathbf{a}_y, d) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}^u, \mathbf{a}_y \in \mathcal{A}, d \in \mathcal{D}^u\}$ is the test set, where $\mathcal{Y}^u$ is the set of labels for unseen classes and $\mathcal{D}^u$ represents the set of unseen domains. In standard ZSL, training and testing images belong to disjoint classes but share the same domain space, *i.e.*, $\mathcal{Y}^s \cap \mathcal{Y}^u \equiv \emptyset$ and $\mathcal{D}^s \equiv \mathcal{D}^u$. In contrast, in the standard DG setting, training and testing im-

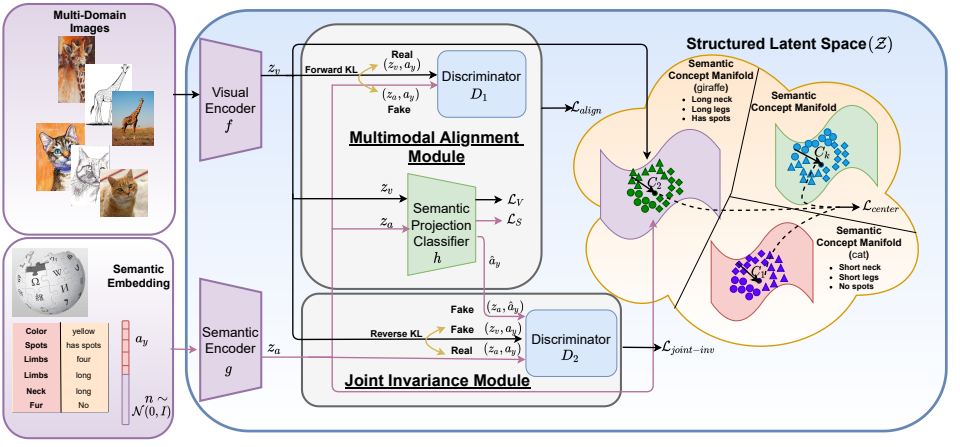

Figure 2: Overall architecture of our approach. The proposed approach comprises a visual encoder f and a semantic encoder g. The multimodal alignment module (Sec. 3.1) aligns the class-specific cues from the visual and semantic latent embeddings ($\mathbf{z}_v$ and $\mathbf{z}_a$) in $\mathcal{Z}$ by employing an alignment loss term $\mathcal{L}_{align}$, The loss term $\mathcal{L}_{center}$ ensures a domain-agnostic class-level partitioning (Sec. 3.2) of $\mathcal{Z}$. Furthermore, the joint invariance module strives to achieve domain invariance (Sec. 3.3) w.r.t the visual-semantic joint distribution by employing $\mathcal{L}_{joint-inv}$, thereby enabling us to generalize to unseen classes in unseen domains.

ages belong to same classes in disjoint domain spaces, *i.e.*, $\mathcal{Y}^s \equiv \mathcal{Y}^u$ and $\mathcal{D}^s \cap \mathcal{D}^u \equiv \emptyset$. Here, our goal is to address the more challenging ZSLDG setting for recognizing unseen classes in unseen domains without having seen these novel classes and domains during training, *i.e.*, $\mathcal{Y}^s \cap \mathcal{Y}^u \equiv \emptyset$ and $\mathcal{D}^s \cap \mathcal{D}^u \equiv \emptyset$.

Overall Framework: The overall architecture of our proposed approach is shown in Fig. 2. The proposed framework comprises a visual encoder f, semantic encoder g, semantic projection classifier h along with discriminators D_1 and D_2. In ZSLDG, the conditional distribution $p(y|\mathbf{x})$ changes since $\mathbf{x}$ comes from different domains, *i.e.*, $p_X(\mathbf{x}|d_i) \neq p_X(\mathbf{x}|d_j)$, $\forall i \neq j$. Our approach mitigates this issue by learning a domain-invariant semantic manifold $\mathcal{Z}$ which is partitioned according to class-level semantic concepts (described in Sec. 3.1 and Sec. 3.2), such that $p(y|\mathbf{z})$ is stable and does not change across domains (where $\mathbf{z} = f(\mathbf{x})$). Furthermore, in order to ensure generalization to unseen classes in unseen domains at test time, our joint invariance module achieves domain invariance w.r.t the visual-semantic joint by employing $\mathcal{L}_{joint-inv}$ (described in Sec. 3.3). This facilitates improved knowledge transfer between class-specific (domain-invariant) visual cues and semantic representations in latent space $\mathcal{Z}$, thereby enhancing generalization to unseen classes in unseen domains at test-time.

3.1 Multimodal Alignment

The multimodal alignment module, learns to project both the visual and semantic representations to a common latent embedding space $\mathcal{Z}$. Let $f(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Z}$ denote a feature extractor, which maps an image $\mathbf{x}$ in the visual space $\mathcal{X}$ to a vector $\mathbf{z}_v$ in the latent embedding space $\mathcal{Z}$. Furthermore, let the function g learn a mapping from semantic space to the latent embedding space, *i.e.*, $g(\mathbf{n}, \mathbf{a}_y) : \mathcal{N} \times \mathcal{A} \rightarrow \mathcal{Z}$ by taking a random Gaussian noise vector $\mathbf{n}$ concatenated with the semantic representation $\mathbf{a}_y$ as input and mapping it to a vector $\mathbf{z}_a$ in $\mathcal{Z}$. Let $D_1 : \mathcal{Z} \times \mathcal{A} \rightarrow \mathsf{R}$ denote a conditional discriminator (conditioned on the semantic embedding $\mathbf{a}_y$) . Then, the multimodal adversarial alignment of the visual and semantic embedding

spaces is achieved by employing a Wasserstein GAN [4], as given by

$$\mathcal{L}_{D_1} = \mathsf{E}[D_1(\mathbf{z}_v, \mathbf{a}_y)] - \mathsf{E}[D_1(\mathbf{z}_a, \mathbf{a}_y)] - \mathit{l}\,\mathsf{E}[(||\tilde{\mathsf{N}}_{\tilde{\mathbf{z}}} D_1(\tilde{\mathbf{z}}, \mathbf{a}_y)||_2 - 1)^2], \tag{1}$$

where $\mathbf{z}_v = f(\mathbf{x})$ and $\mathbf{z}_a = g(\mathbf{n}, \mathbf{a}_y)$ are the latent embeddings from the visual and semantic spaces, respectively. Here, l is a weighting coefficient, while $\tilde{\mathbf{z}} = h\mathbf{z}_v + (1-h)\mathbf{z}_a$ with $h \sim U(0,1)$ represents a convex combination of $\mathbf{z}_v$ and $\mathbf{z}_a$. Eq. 1 is equivalent to minimizing the (forward) Kullback-Leibler (KL) divergence between the visual and semantic latent embeddings, *i.e.*, $KL[(\mathbf{z}_v, \mathbf{a}_y)||(\mathbf{z}_a, \mathbf{a}_y)]$. Furthermore, to enhance the discriminability of learned latent embeddings, we employ a compatibility based classifier using a semantic projection function $h : \mathcal{Z} \rightarrow \mathcal{A}$ for constraining the latent embeddings ($\mathbf{z}_v$ and $\mathbf{z}_a$) to map back to their corresponding semantic representations $\mathbf{a}_y$, given by,

$$\mathcal{L}_V(\mathbf{z}_v, \mathbf{a}_y) = -\mathsf{E}(\log \frac{\exp(\langle h(\mathbf{z}_v), \mathbf{a}_y \rangle)}{\mathsf{S}_{\mathbf{y} \in Y^s} \exp(\langle h(\mathbf{z}_v), \mathbf{a}_y \rangle)}), \quad \mathcal{L}_S(\mathbf{z}_a, \mathbf{a}_y) = -\mathsf{E}(\log \frac{\exp(\langle h(\mathbf{z}_a), \mathbf{a}_y \rangle)}{\mathsf{S}_{\mathbf{y} \in Y^s} \exp(\langle h(\mathbf{z}_a), \mathbf{a}_y \rangle)}). \tag{2}$$

Here, $\langle \cdot, \cdot \rangle$ measures the similarity between its inputs, computed as the dot product between them. The cyclic projection of mapping from visual/semantic space to a latent space and then back to the semantic space minimizes the information loss and enhances the latent embedding discriminabilty. We employ the multimodal alignment loss term ($\mathcal{L}_{align}$) to learn the visual and semantic encoders along with the semantic projection classifier, given by

$$\mathcal{L}_{align} = \mathsf{E}[D_1(\mathbf{z}_v, \mathbf{a}_y)] - \mathsf{E}[D_1(\mathbf{z}_a, \mathbf{a}_y)] + \mathcal{L}_V(\mathbf{z}_v, \mathbf{a}_y) + \mathcal{L}_S(\mathbf{z}_a, \mathbf{a}_y)]. \tag{3}$$

3.2 Structured Partitioning

While the multimodal alignment module aligns the visual and corresponding semantic embeddings in the latent space, it does not learn a domain-agnostic latent space, which is partitioned according to the semantic concepts that relate to the different classes. In order to achieve a structured and domain-invariant latent space, we propose to cluster the latent embeddings based on class-level (domain-invariant) semantic concepts across different domains. The latent space is then conceptually structured, since the visual latent embeddings $\mathbf{z}_v$ and semantic latent embeddings $\mathbf{z}_a$ of a class are clustered together. To this end, we adopt the center loss [44] in a multimodal setting. Formally, we first randomly initialise S centers, *i.e.*, $\{\mathbf{c}_j | j = 1, \ldots, S\}$ for each of the seen classes in the training set and compute the loss, $\mathcal{L}_{center}$ due to each class y present in a mini-batch. Then, for every class y that is present in a mini-batch, the center update $\mathsf{D}\mathbf{c}_y$ is computed for incrementing the corresponding center $\mathbf{c}_y$. The loss $\mathcal{L}_{center}$ and update $\mathsf{D}\mathbf{c}_y$ are given by:

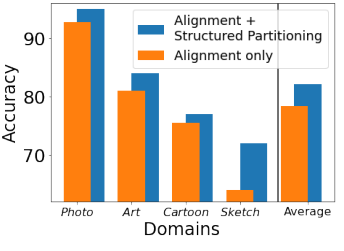

Figure 3: Impact of our structured partitioning for the DG task on PACS [20]. Compared to multimodal alignment alone (orange bars), additionally partitioning the latent space according to the semantic concepts along with multimodal alignment provides notable performance gains (blue bars), especially on the most difficult unseen domain, *i.e.*, *sketch*.

$$\mathcal{L}_{center} = d[\mathsf{E}(||\mathbf{z}_v - \mathbf{c}_y||_2^2) + \mathsf{E}(||\mathbf{z}_a - \mathbf{c}_y||_2^2)]; \qquad \mathsf{D}\mathbf{c}_y = \mathsf{E}[\mathbf{c}_y - \mathbf{z}_v] + \mathsf{E}[\mathbf{c}_y - \mathbf{z}_a]. \tag{4}$$

Here, $\mathbf{c}_y$ denotes the center of class label y in the latent space, while $\mathbf{z}_v$ and $\mathbf{z}_a$ correspond to the visual and semantic embeddings of class y, and d is weighing factor for center loss. Consequently, the intra-class and inter-domain variances for each class get minimized, resulting

in a structured and domain-agnostic latent space. Furthermore, since both the visual and semantic latent representations of a class are clustered together, the latent space is partitioned based on class-level semantic concepts.

In order to validate our hypotheses that a domain-agnostic structured latent space helps to stabilize $p(y|\mathbf{z})$ and generalize to new domains, we conduct an experiment as a proof of concept. Fig. 3 presents a comparison for the standard domain generalization (DG) setting on the PACS dataset [20] using ResNet-18 backbone. We see that structuring the latent space (blue bars) provides performance gains on all domains and enhances the average gain, compared to employing multimodal alignment alone (orange bars). The highest gain is achieved for the most difficult *sketch* domain that has a large domain shift from the source domains (*photo, art, cartoon*), demonstrating the advantage of our domain-agnostic partitioning.

3.3 Joint Invariance Module

As discussed above, the multimodal alignment and conceptual partitioning result in a structured and domain-agnostic latent embedding space that disentangles semantic and domain-specific information. Such a disentanglement of semantic and domain-specific information is sufficient for standard domain-generalization setting where images during training and testing come from same categories. However in our ZSLDG setting, the disentanglement may not hold for unseen semantic categories during testing, as previously found in [26]. In order to address this issue and enable generalization to unseen classes in unseen domains, we propose to learn the domain-invariance w.r.t the joint distribution of visual and semantic representations of a class. Formally, any given image $\mathbf{x}$ comprises of a class-specific content $\mathbf{C}_y$ and a domain-specific transformation $T_i(\cdot)$ which depicts the class y in that particular domain d_i. Thus, each image $\mathbf{x} \in \mathcal{X}$ belonging to domain d_i can be represented as $\mathbf{x} = T_i(\mathbf{C}_y)$. In order to enable generalization to unseen class in unseen domains, we propose to match the visual-semantic joint distribution $p(T_i(\mathbf{C}_y), \mathbf{a}_y)$ under different domain transformations $T_i(\cdot)$ in order to disentangle the domain-specific information. Since the semantic space is shared between seen and unseen classes, learning domain invariance w.r.t the joint distribution of visual and semantic representations of a class, *i.e.*, $p(f(T_i(\mathbf{C}_y)), \mathbf{a}_y)$ or $p(f(\mathbf{x}), \mathbf{a}_y)$ enables us to enhance generalization.

Specifically, we aim to match the visual-semantic joint distribution from the visual encoder $(\mathbf{z}_v, \mathbf{a}_y)$, semantic encoder $(\mathbf{z}_a, \mathbf{a}_y)$ and projection classifier $(\mathbf{z}_a, \hat{\mathbf{a}}_y)$. To this end, we employ a triple adversarial loss, to stabilize the visual-semantic joint distribution across different domains. This also enhances visual-semantic interaction for learning class-specific discriminative features in the visual and semantic embedding spaces. This is achieved by employing a discriminator $D_2 : \mathcal{Z} \times \mathcal{A} \rightarrow \mathsf{R}$ and optimizing:

$$\mathcal{L}_{D_2} = \mathsf{E}[D_2(\mathbf{z}_a, \mathbf{a}_y)] - a\,\mathsf{E}[D_2(\mathbf{z}_a, \hat{\mathbf{a}}_y)] - b\,\mathsf{E}[D_2(\mathbf{z}_v, \mathbf{a}_y)] - l\,\mathsf{E}[(||\tilde{\mathsf{N}}_{\tilde{\mathbf{z}}} D_2(\tilde{\mathbf{z}}, \tilde{\mathbf{a}}_y), \tilde{\mathsf{N}}_{\tilde{\mathbf{a}}} D_2(\tilde{\mathbf{z}}, \tilde{\mathbf{a}}_y)||_2 - 1)^2] \quad (5)$$

Here, $\hat{\mathbf{a}}_y = h(\mathbf{z}_a)$ is output from projection classifier h, which represents the projection of the latent embedding $\mathbf{z}_a$ onto the semantic space $\mathcal{A}$. Also, $\tilde{\mathbf{z}} = h\mathbf{z}_a + (1-h)(a\mathbf{z}_a + b\mathbf{z}_v)$ and $\tilde{\mathbf{a}}_y = h\mathbf{a}_y + (1-h)(a\hat{\mathbf{a}}_y + b\mathbf{a}_y)$ with $b = 1-a$ and $h \sim U(0,1)$. Additionally, l is a weighting coefficient. Note that D_2 is different from the vanilla discriminator D_1 and has a triple adversarial formulation [19]. Firstly, by incorporating the projection classifier output $\hat{\mathbf{a}}_y$, it enables to jointly train the visual encoder f, semantic encoder g and projection classifier h while imposing domain-invariance. In addition, we design Eq. 5 to treat $(\mathbf{z}_a, \mathbf{a}_y)$ as real samples and $(\mathbf{z}_v, \mathbf{a}_y)$, $(\mathbf{z}_a, \hat{\mathbf{a}}_y)$ as fake samples. This acts as a minimizer of the reverse

KL divergence *i.e.*, $KL[(z_a, a_y)||(z_v, a_y)]$ [34] (in contrast to D_1 that minimizes forward KL as described in Sec. 3.1) between the visual and semantic spaces. We find that this leads to better generalization by alleviating the mode collapse issue, and thus enables our model to capture multiple modes of the data distribution [34]. Next, the semantic projector classifier h is updated to minimize:

$$\mathcal{L}_{cls} = -a\,\mathsf{E}[p_h(y|\mathbf{z}_a)D_2(\mathbf{z}_a, \hat{\mathbf{a}}_y)] + g[\mathcal{L}_V(\mathbf{z}_v, \mathbf{a}_y) + \mathcal{L}_S(\mathbf{z}_a, \mathbf{a}_y)], \tag{6}$$

where $p_h(y|\mathbf{z}_a)$ is the probability distribution after taking softmax of semantic projection classifier h, output logits. Weighting the D_2 output with the class probabilities helps in achieving stable training [19]. Finally, we update the visual and semantic encoders (f and g) to minimize discrepancy between the embeddings ($\mathbf{z}_a$ and $\mathbf{z}_v$) in the latent space, given by:

$$\mathcal{L}_{gen} = \mathsf{E}[D_2(\mathbf{z}_a, \mathbf{a}_y)] - b\,\mathsf{E}[D_2(\mathbf{z}_v, \mathbf{a}_y)]. \tag{7}$$

Then, the joint invariance loss term $\mathcal{L}_{joint-inv}$ is defined as $\mathcal{L}_{joint-inv} = \mathcal{L}_{cls} + \mathcal{L}_{gen}$. Consequently, the adversarial loss terms in Eq. 5 and $\mathcal{L}_{joint-inv}$ together enable us to jointly train f, g, h and learn a domain-invariant space, which can generalize to unseen domains and classes at test time, by capturing class-specific discriminative visual-semantic relationships across domains.

3.4 Training and Inference

Training: In a single training iteration, we update the discriminators D_1 and D_2 to maximize the losses in Eq. 1 and 5. We update the discriminators 5 times for every update of the rest of the functions (f, g, h), as in WGAN [14]. Following this, the parameters q_f, q_g, q_h, q_c corresponding to f, g, h and class centers, respectively, are updated to minimize:

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \mathcal{L}_{center} + \mathcal{L}_{joint-inv}. \tag{8}$$

Inference: A test image $\mathbf{x}_t$ from a unseen domain and class (in $\mathcal{D}^u$ and $\mathcal{Y}^u$) is projected by encoder f to obtain the corresponding latent embedding $\mathbf{z}_t = f(\mathbf{x}_t)$. The semantic projection classifier h computes pairwise similarities between $\mathbf{z}_t$ and the unseen class embeddings $\mathbf{a}_y$, where $y \in \mathcal{Y}^u$. These similarity scores are converted to class probabilities to obtain the final prediction $\hat{y}$, given by $\hat{y} = \arg\max_{y \in \mathcal{Y}^u} P(y|\mathbf{x}_t; \mathsf{F})$.

4 Experiments

Datasets: We evaluate our method on the DomainNet and DomainNet-LS benchmarks for the task of ZSLDG, as in [26]. **DomainNet [36]:** It is a large-scale dataset and is currently the only benchmark dataset for the ZSLDG setting [26]. It consists of nearly 0.6 million images from 345 categories in 6 domains: *painting*, *clipart*, *sketch*, *infograph*, *quickdraw* and *real*. For the task of ZSLDG, we follow the same training/validation/testing splits along with the training and evaluation protocol described in [26]. In particular, 45 out of 345 are fixed as unseen classes and training is performed using only the remaining seen class images. Among the 6 domains in DomainNet, the seen class images from 5 domains are provided during training, and the model is evaluated on the 45 unseen classes in the held-out (unseen) domain. We repeat experiments with each of the domains as the unseen domain. Following [26], the *real* domain is never held out since a ResNet-50 backbone, pre-trained on ImageNet [8], is employed. Average per-class accuracy is used as the performance metric for evaluation on the held-out domain [26, 47]. Similarly, we use the *word2vec* [29] representations as the

Table 1: State-of-the-art comparison for the task of ZSLDG on the DomainNet benchmark using ResNet-50 backbone [26]. For a fair comparison, all reported results employ the same backbone, metrics, protocol and splits, as described in [26]. Best results are in bold.

Method		AVG	Target Domain				
DG	ZSL		*painting*	*infograph*	*quickdraw*	*sketch*	*clipart*
-	DEVISE [11]	14.4	17.6	11.7	6.1	16.7	20.1
	ALE [3]	16.2	20.2	12.7	6.8	18.5	22.7
	SPNet [46]	19.4	23.8	16.9	8.2	21.8	26.0
DANN [12]	DEVISE [11]	13.9	16.4	10.4	7.1	15.1	20.5
	ALE [3]	15.7	19.7	12.5	7.4	17.9	21.2
	SPNet [46]	19.1	24.1	15.8	8.4	21.3	25.9
EpiFCR [22]	DEVISE [11]	15.9	19.3	13.9	7.3	17.2	21.6
	ALE [3]	17.5	21.4	14.1	7.8	20.9	23.2
	SPNet [46]	20.0	24.6	16.7	9.2	23.2	26.4
CuMix (Mixup-img-only)		19.2	24.4	16.3	8.7	21.7	25.2
CuMix (Mixup-two-level)		19.9	25.3	17	8.8	21.9	26.6
CuMix [26]		20.7±0.34	25.5±0.40	17.8±0.20	9.9±0.33	22.6±0.30	27.6±0.50
Ours		**21.9±0.29**	**26.6±0.30**	**18.4±0.40**	**11.5±0.18**	**25.2±0.3**	**27.8±0.28**

Table 2: State-of-the-art comparison using standard accuracy for the task of ZSLDG on DomainNet using ResNet-50 backbone. CuMix [26] results are reproduced using the public code. For a fair comparison, we employ the same backbone, protocol and splits, as in [26]. Best results are in bold.

Method	AVG	Target Domain				
		painting	*infograph*	*quickdraw*	*sketch*	*clipart*
CuMix [26]	21.5	27.6	16.3	9.7	25.9	27.8
Ours	**22.7**	**28.8**	**17.6**	**11.5**	**26.3**	**29.1**

semantic information for inter-relating seen and unseen classes, as in [26]. The **DomainNet-LS** benchmark is more challenging, where the source domains during training are limited to *real* and *painting* only, whereas testing is conducted on the remaining four unseen domains. Since only two source domains are used in training, it is more challenging to learn domain-invariance and generalize at test-time.

Implementation Details: The semantic encoder g, semantic projection classifier h and the discriminators D_1 and D_2 are implemented as fully connected (FC) networks. The semantic encoder g is a two-layer FC network with hidden layer of size 4,096. Its output dimension of 2,048 matches the output dimension of the visual encoder f. While the discriminators D_1 and D_2 are also two-layer FC networks with hidden layers of size 4,096, the semantic projection classifier h is a single-layer FC network. Leaky ReLU activation is used everywhere, except at the output of g, which has a ReLU non-linearity. The visual encoder f is the standard ResNet-50 [16] backbone, as in [26]. The semantic projection classifier with loss L_V is trained first for a few iterations as a warm-up followed by an end-to-end training for all modules. We use the Adam optimizer [17] with a learning rate of 10^{-4}. We do not use any special scheduling procedure for the learning rate. We set $d = 0.01$ and $a = 0.5$ for all domains. Similarly, we find that best results are obtained for $g \in [5,7]$ for all domains.

4.1 Results: Comparison with State-of-the-art

Results on DomainNet: Tab. 1 shows the comparison of our proposed framework with state-of-the-art methods and all baselines, as established in [26], on the ZSLDG task. Following protocol established in [26], we report Average per-class accuracy as the performance metric for evaluation on the held-out domain. We first report the performance of standalone ZSL approaches such as DEVISE [11], ALE [3] and SPNet [46] on the ZSLDG task, followed by the performance achieved by coupling these ZSL approaches with standard DG approaches like DANN [12] and EpiFCR [22]. It is worth noting that coupling the standalone ZSL methods with DANN achieves lower performance than the ZSL method alone in the case of ZSLDG, since standard domain alignment methods have been shown to be ineffective on the DomainNet dataset, leading to negative transfer in some cases [36]. Fur-

Table 3: Results on DomainNet-LS with only *real* and *painting* as source domains and ResNet-50 backbone, following protocol in [26]. Best results are in bold.

Model	AVG	*quickdraw*	*sketch*	*infograph*	*clipart*
SPNet[ ]	14.4	4.8	17.3	14.1	21.5
Epi-FCR[ ]+SPNet[ ]	15.4	5.6	18.7	14.9	22.5
CuMix(MixUp-img-only)[ ]	14.3	4.8	17.3	14.0	21.2
CuMix(MixUp-two-level)[ ]	15.8	4.9	19.1	16.5	22.7
CuMix (reverse)[ ]	15.4	4.8	18.2	15.8	22.9
CuMix[ ]	16.5	5.5	19.7	**17.1**	23.7
Ours	**16.9**	**7.2**	**20.5**	16	**24**

Table 4: Ablation study for different components of our framework on DomainNet dataset for ZSLDG setting. Best results are in bold.

Model	AVG	*painting*	*infograph*	*quickdraw*	*sketch*	*clipart*
M1: L_{align}	18.5	22.6	16.2	9.6	20.8	23.7
M2: M1 + L_{center}	20.5	25.4	16.9	9.8	24.0	26.4
M3: M2 + $L_{joint-inv}$	**21.9**	**26.6**	**18.4**	**11.5**	**25.2**	**27.8**

thermore, as noted by [26], coupling EpiFCR (a standalone DG method) with the standalone ZSL approaches is not straightforward, since it requires careful adaptation that includes restructuring of the loss terms. In particular, the approach of EpiFCR+SPNet achieves an average accuracy (AVG) of 20.0 over different target domains. The recently introduced CuMix [26] approach that targets ZSLDG, employs a curriculum-based mixing policy to generate increasingly complex training samples by mixing up multiple seen domains and categories available during training. The current state-of-the-art CuMix improves ZSLDG performance over EpiFCR+SPNet, achieving an average accuracy of 20.7 across the target domains. Our approach outperforms CuMix with an absolute average gain of 1.2% across domains ($\sim$ 5.8% relative increase) and achieves an average accuracy of 21.9 across the five target domains, setting a new state of the art. Furthermore, our method achieves consistent gains over CuMix on each of the target domains.

In addition, to facilitate a holistic evaluation, Tab. 2 shows the performance comparison of our proposed method with the best existing CuMix [26] approach using the standard accuracy metric. Our approach achieves consistent gains over all domains and improves over CuMix when using standard accuracy as the metric, with an absolute average gain of 1.2% across domains ($\sim$ 5.6% relative gain), including significant relative gains of about 18% for harder domains like *quickdraw*. These results show that the proposed method obtains favorable performance over existing approaches on different evaluation metrics.

Results on DomainNet-LS: Tab. 3 shows the performance comparison on the DomainNet-LS benchmark. The SPNet [46] (for standard ZSL) achieves an average accuracy of 14.4, while its integration with EpiFCR [22] (a standard DG approach) improves the performance to 15.4. The best existing CuMix [26] approach for ZSLDG achieves 16.5 as the average accuracy across the unseen domains. Despite the limited information available during training (and higher domain shift at test time), our approach improves over CuMix by achieving an average accuracy of 16.9 (2.4% relative gain), thereby showing better generalization.

4.2 Ablation Study

We perform an ablation study to understand the efficacy of each component in our proposed method for the ZSLDG task. Tab. 4 shows the performance gains achieved (on DomainNet [36]) by integrating one contribution at a time, in our approach, as below:

- The model learned by employing our multimodal alignment loss term $\mathcal{L}_{align}$ alone (detailed in Sec. 3.1) is denoted as M1
- Similarly, M2 denotes the model learned by integrating $\mathcal{L}_{align}$ with our loss term $\mathcal{L}_{center}$, which achieves a structured latent space (Sec. 3.2).
- M3 denotes our overall framework, which is learned by integrating our joint invariance loss term $\mathcal{L}_{joint-inv}$ (Sec. 3.3) with $\mathcal{L}_{align}$ and $\mathcal{L}_{center}$.

The M1 model, which performs multimodal adversarial alignment achieves an average accuracy (denoted as AVG in Tab. 4) of 18.5 across the target domains. Learning a structured latent embedding space along with the multimodal alignment enables the M2 model to achieve

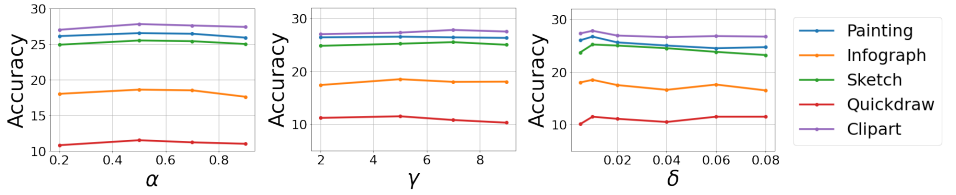

Figure 4: Robustness analysis of our proposed approach w.r.t the parameters a, g and d on different domains (right corner) in DomainNet. It can be seen that our approach exhibits fairly stable trend with variations in a, g and d.

an average gain of 2.0 over M1 on the target domains. We note that the gains in M2 due to the integration of $\mathcal{L}_{center}$ with $\mathcal{L}_{align}$ are considerably high on the easier target domains (*clipart*, *painting* and *sketch*). This suggests that $\mathcal{L}_{center}$ is able to achieve an improved structuring of the latent embedding space. Our overall framework (M3) obtains the best results by achieving an average accuracy of 21.9 on the five target domains. Since M3 additionally involves learning the domain invariance w.r.t the visual-semantic joint by employing $\mathcal{L}_{joint-inv}$, it aids in improving ZSLDG performances on harder target domains such as *quickdraw* and *infograph*. These results clearly indicate that along with the multimodal alignment ($\mathcal{L}_{align}$), structuring the latent space ($\mathcal{L}_{center}$) and learning the domain invariance w.r.t the visual-semantic joint ($\mathcal{L}_{joint-inv}$) are important for recognizing unseen classes in unseen domains.

4.3 Effect of Hyperparameter Variations

Fig. 4 shows the performance variation of our method for different choices of hyperparameters. Our framework consists mainly of three important hyperparameters, that is, a , g and d (as defined in Eq. 5, 6 and 4, respectively). Note that $b = 1 - a$, as defined in Sec 3.3. We notice that (Fig. 4, left) our approach achieves the best performance at a=0.5 for all domains. This implies that it is important to match joint distributions $(\mathbf{z}_a, \hat{\mathbf{a}}_y)$ from the semantic projection classifier and $(\mathbf{z}_v, \mathbf{a}_y)$ from the visual encoder; with the joint distribution $(\mathbf{z}_a, \mathbf{a}_y)$ from semantic encoder in Eq. 5. In addition, we notice that (Fig. 4, middle) the performance improves as g increases likely due to enhanced discriminability of embeddings in latent space $\mathcal{Z}$. The best results are obtained for $g \in [5,7]$, after which the performance saturates and tends to slowly drop. Furthermore, we observe that (Fig. 4, right) $d = 0.01$ obtains the best performance on all domains. We also notice that for both relatively easier and harder domains, the trend of performance is consistent and fairly stable with change in hyperparameters. *Additional ablations and analysis are presented in the supplementary*.

5 Conclusions

We propose a novel approach to address the challenging problem of recognizing unseen classes in unseen domains (ZSLDG). Our method learns a domain-agnostic structured latent embedding space which is achieved by employing a multimodal alignment loss term that aligns the visual and semantic spaces, a center loss term that separates different classes in the latent space and a joint invariance term that aids in handling new classes from unseen domains. Our experiments and ablation studies on challenging benchmarks (DomainNet, DomainNet-LS) show the superiority of our approach over existing methods. Future directions include leveraging self-supervision to obtain domain-invariant features and tackle dynamic changes in the label space of categories.

Acknowledgement: This work has been partly supported by the funding received from Academy of Finland grant 329268 "Movie Making Finland: Finnish fiction films as audio-visual big data, 1907–2017"

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. *CVPR*, 2015.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label embedding for image classification. *TPAMI*, 2016.

[3] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826, 2013.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *ICML*, 2017.

[5] Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018.

[6] Gaurav Bhatt, Shivam Chandhok, and Vineeth N. Balasubramanian. Learn from anywhere: Rethinking generalized zero-shot learning with limited supervision. *ArXiv*, abs/2107.04952, 2021.

[7] Shivam Chandhok and V. Balasubramanian. Two-level adversarial visual-semantic coupling for generalized zero-shot learning. *WACV*, 2021.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[9] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *ICLR Workshop*, 2015.

[10] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. *ECCV*, 2018.

[11] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, pages 2121–2129, 2013.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[13] Muhammad Ghifary, W. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. *ICCV*, pages 2551–2559, 2015.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *NeurIPS*, 2017.

[15] Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost van de Weijer. Generative multi-label zero-shot learning. *arXiv preprint arXiv:2101.11606*, 2021.

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[18] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into crossspace mapping for zero-shot learning. *ACL*, 2015.

[19] Chongxuan Li, T. Xu, J. Zhu, and B. Zhang. Triple generative adversarial nets. In *NeurIPS*, 2017.

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. *ICCV*, pages 5543–5551, 2017.

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[22] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019.

[23] Haoliang Li, Sinno Jialin Pan, S. Wang, and A. Kot. Domain generalization with adversarial feature learning. *CVPR*, pages 5400–5409, 2018.

[24] J. Li, Mengmeng Jing, Ke Lu, Z. Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. *CVPR*, pages 7394–7403, 2019.

[25] Y. Li, X. Tian, Mingming Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018.

[26] Massimiliano Mancini, Zeynep Akata, E. Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *ECCV*, 2020.

[27] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In *CVPR*, 2019.

[28] Udit Maniyar, K. J. Joseph, A. Deshmukh, Ü. Dogan, and V. Balasubramanian. Zero-shot domain generalization. *ArXiv*, abs/2008.07443, 2020.

[29] Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[30] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. *CVPR Workshop*, 2018.

[31] Krikamol Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. *ArXiv*, abs/1301.2115, 2013.

[32] Sanath Narayan, A. Gupta, F. Khan, Cees G. M. Snoek, and L. Shao. Latent embedding feedback and discriminative features for zero-shot classification. *ArXiv*, abs/2003.07833, 2020.

[33] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *ICCV*, pages 8731–8740, October 2021.

[34] T. Nguyen, Trung Le, H. Vu, and Dinh Q. Phung. Dual discriminator generative adversarial nets. In *NeurIPS*, 2017.

[35] Jian Ni, Shanghang Zhang, and Haiyong Xie. Dual adversarial semantics-consistent network for generalized zero-shot learning. *NeurIPS*, 2019.

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. *ICCV*, pages 1406–1415, 2019.

[37] B. Romera-Paredes and P. H. Torr. An embarrassingly simple approach to zero-shot learning. *ICML*, 2015.

[38] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. *CVPR*, 2019.

[39] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, S. Chaudhuri, P. Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *ArXiv*, abs/1804.10745, 2018.

[40] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 2013.

[41] Y.-H. H. Tsai, L.-K. Huang, and R. Salakhutdinov. Learning robust visual-semantic embeddings. *ICCV*, 2017.

[42] Riccardo Volpi, Hongseok Namkoong, O. Sener, John C. Duchi, Vittorio Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018.

[43] M. R. Vyas, Hemanth Venkateswara, and S. Panchanathan. Leveraging seen and unseen semantic relationships for generative zero-shot learning. In *ECCV*, 2020.

[44] Y. Wen, Kaipeng Zhang, Z. Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[45] Yongqin Xian, Tobias Lorenz, Bernt Schiele, , and Zeynep Akata. Feature generating networks for zero-shot learning. *CVPR*, 2018.

[46] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, pages 8256–8265, 2019.

[47] Yongqin Xian, Christoph H. Lampert, B. Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 41:2251–2265, 2019.

[48] Yongqin Xian, Saurabh Sharma, B. Schiele, and Zeynep Akata. F-vaegan-d2: A feature generating framework for any-shot learning. *CVPR*, pages 10267–10276, 2019.

[49] Zheng Xu, W. Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014.

[50] P. Yang and Wei Gao. Multi-view discriminant transfer learning. In *IJCAI*, 2013.