TIDE: Training Locally Interpretable Domain Generalization Models Enables Test-time Correction

Aishwarya Agarwal[1,2*] *Srikrishna Karanam*[2†] *Vineet Gandhi*[1‡]

[1]CVIT, Kohli Centre for Intelligent Systems, IIIT Hyderabad, India

[2]Adobe Research, Bengaluru, India

Abstract

We consider the problem of single-source domain generalization. Existing methods typically rely on extensive augmentations to synthetically cover diverse domains during training. However, they struggle with semantic shifts (e.g., background and viewpoint changes), as they often learn global features instead of local concepts that tend to be domain invariant. To address this gap, we propose an approach that compels models to leverage such local concepts during prediction. Given no suitable dataset with per-class concepts and localization maps exists, we first develop a novel pipeline to generate annotations by exploiting the rich features of diffusion and large-language models. Our next innovation is TIDE, a novel training scheme with a concept saliency alignment loss that ensures model focus on the right per-concept regions and a local concept contrastive loss that promotes learning domain-invariant concept representations. This not only gives a robust model but also can be visually interpreted using the predicted concept saliency maps. Given these maps at test time, our final contribution is a new correction algorithm that uses the corresponding local concept representations to iteratively refine the prediction until it aligns with prototypical concept representations that we store at the end of model training. We evaluate our approach extensively on four standard DG benchmark datasets and substantially outperform the current state-of-the-art (12% improvement on average) while also demonstrating that our predictions can be visually interpreted.

1. Introduction

Enhancing deep neural networks to generalize to out-of-distribution samples remains a core challenge in machine learning and computer vision research, as real-world test data often diverges significantly from the training distribution [31]. The challenges compound when obtaining labeled samples from the target domain is expensive or unfeasible, hindering application of semi-supervised learning or domain adaptation [25, 32, 49]. The problem of Domain Generalization (DG) [30, 44, 54, 61, 63] represents a promising avenue for developing techniques that capture domain-invariant patterns and improve performance on out-of-distribution samples. In this paper, we focus on Single Source Domain Generalization (SSDG), where a model is trained on data from a single domain and aims to generalize well to unseen domains [40, 51]. It represents the most strict form of DG, as the model must extract domain-invariant features from a single, often limited perspective, without exposure to the variation present across multiple domains.

Most previous work on SSDG [5, 7, 8, 19, 20, 22, 59, 63] relies extensively on data augmentations to support the learning of domain-generalized features. The premise is that constructing an extensive repertoire of augmentations synthesizes instances encompassing a wide spectrum of

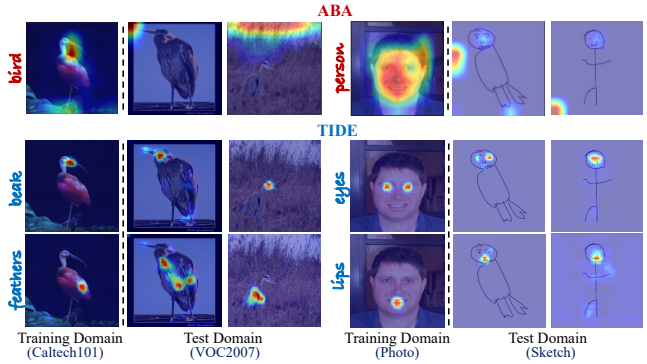

Figure 1. Samples from VLCS (left) and PACS dataset (right) across domain shifts, corresponding to bird and person class. First row displays GradCAM maps [43] for ABA's class predictions. We observe that model attention of ABA [5] falters across domain shifts. Second and third row display the concept specific GradCAM maps from TIDE. We posit that accurate concept learning and localization facilitates DG.

*aishwarya.agarwal@research.iiit.ac.in, aishagar@adobe.com

†skaranam@adobe.com

‡vgandhi@iiit.ac.in

human-recognizable domains. However, accounting for all conceivable real-world augmentations presents an immense challenge. These models have shown reasonable success in addressing variations in style and texture [45, 64]; however, their performance remains modest when faced with more semantic domain shifts, such as changes in background and viewpoint e.g., in the VLCS dataset [41, 48].

Consider the results of the state-of-the-art ABA [5] model on the two examples in Figure 1. In the first example (left), domain shifts manifest as variations in background and viewpoint. In the second example (right), the training data comprises photos of human faces, whereas the test data consists of skeletal sketches. In both cases, the domain shifts extend beyond style or texture variations, with ABA failing to correctly classify the samples. Class-level saliency maps [43] reveal that this misclassification stems from inadequate focus on *critical local concepts* under domain shifts, such as `beak` and `feathers` for birds, or `eyes` and `lips` for persons. Such local concepts are integral to class definition and remain invariant across domains, and a robust DG model must adeptly capture these stable features. We posit that prior efforts learn global features per class; and if the model fails to learn the correct feature set in the training domain, its generalization performance is compromised as noted from Figure 1 above. The inconsistency in concept localization further exacerbates the interpretability and explainability of these models.

We adopt an alternative approach, wherein, rather than attempting to encompass all potential augmentations, we compel the model to leverage *essential class-specific concepts* for classification. Our key idea is to force the model to attend to these concepts during training. The primary hurdle in this path is lack of annotated data specifying relevant class-specific concepts along with their corresponding localizations. To this end, our first contribution is a novel pipeline that harnesses large language models (LLMs) and diffusion models (DMs) to identify key concepts and generate concept-level saliency maps in a *scalable*, *automated fashion*. We demonstrate that DMs, via cross-attention layers, can generate concept-level saliency maps for generated images with notable granularity. Given these maps for a single synthesized image within a class, we leverage the rich feature space of DMs to transfer them to images across diverse domains in DG benchmarks [47].

We subsequently introduce TIDE, our second contribution which employs cross-entropy losses over class and concept labels along with a novel pair of loss functions: a *concept saliency alignment loss* that ensures the model attends to the correct regions for each local concept, and a *local concept contrastive loss* that promotes learning domain invariant features for these regions. As shown in Figure 1, TIDE consistently attends to local concepts such as `beak` and `eyes` in the training domain as well as across substantial domain shifts. This precise localization significantly bolsters SSDG performance while enabling the generation of visual explanations for model decisions.

Marking a major leap in model interpretability and performance gains, our third contribution is to demonstrate how our model's concept-level localization can be effectively leveraged for *performance verification* and *test-time correction*. To this end, we introduce the notion of *local concept signatures*: prototypical features derived by pooling concept-level features across training samples, guided by corresponding saliency maps. If the concept features associated with the predicted class label do not align with their signatures, it signals the use of incorrect features for prediction. Consequently, we employ iterative refinement through concept saliency masking until concept predictions align with their corresponding signatures. More formally, our paper makes following contributions:

- We propose a novel synthetic annotation pipeline to automatically identify key per-class local concepts and their localization maps, and transfer them to real images found in DG benchmarks.
- We propose a novel approach, TIDE, to train locally interpretable models with a novel concept saliency alignment loss that ensures accurate concept localization and a novel local concept contrastive loss that ensures domain-invariant features.
- We show TIDE enables correcting model prediction at test time using the predicted concept saliency maps as part of a novel iterative attention refinement strategy.
- We report extensive results on four standard DG benchmarks, outperforming the current state-of-the-art by a significant 12% on average.

2. Related Works

Multi Source domain generalization (MSDG) methods assume access to multiple source domains and domain-specific knowledge during training and have proposed ways to learn multi-source domain-invariant features [10, 15, 34, 35, 53], utilize domain labels to capture domain shifts [11, 35, 56], and design target-specific augmentation strategies [14, 16, 23]. However, these assumptions are not practical for real-world applications [44, 57] and we instead focus on the more challenging SSDG setting, where only one source domain and no prior target knowledge is available.

Most SSDG approaches have used augmentations to improve DG [5, 7, 8, 19, 20, 22, 59, 63]. While [59] also uses diffusion models (DMs) [42] for data augmentation, our approach differs by leveraging the rich feature space of DMs primarily for offline saliency map annotation instead of augmentation. More recent work in SSDG [4, 28, 29] has also turned to approaches based on domain-specific prompting and causal inference instead of data augmentations. However, these methods learn global features, limiting invari-

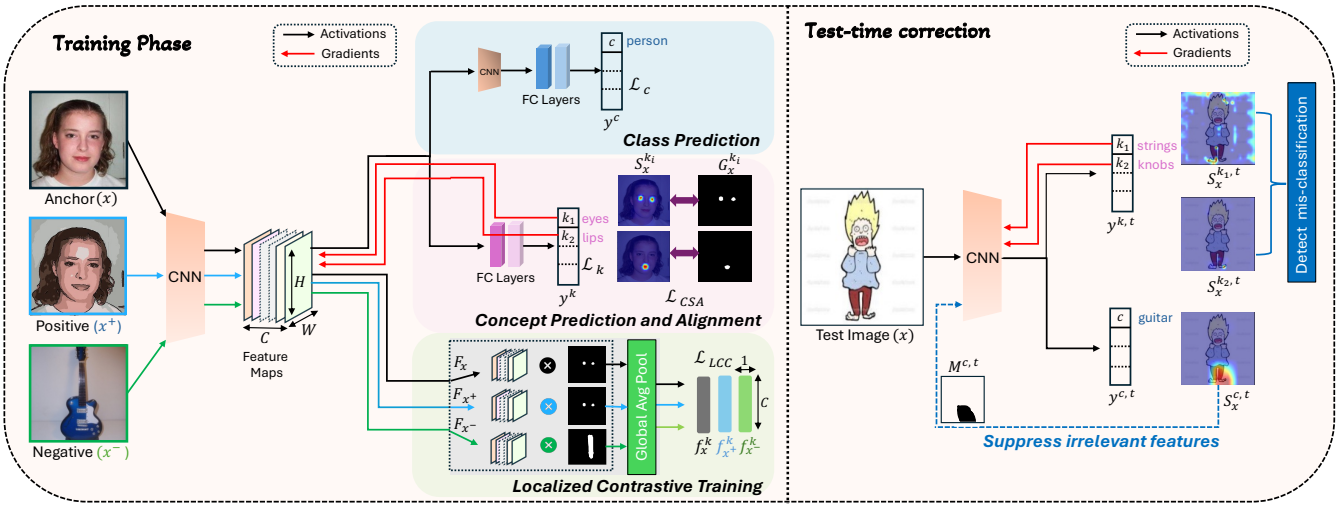

Figure 2. The TIDE pipeline: Left—Training on a single domain with cross-entropy losses for class ($\mathcal{L}_c$) and concept labels ($\mathcal{L}_k$), alongside Concept Saliency Alignment ($\mathcal{L}_{CSA}$) and Local Concept Contrastive losses ($\mathcal{L}_{LCC}$). Right—Test-time correction strategy applied in TIDE.

ance across unseen domains. Finally, these methods also do not provide any signals to interpret model predictions.

On the other hand, while methods like Concept Bottleneck Models (CBMs) [24, 33] also learn local concepts, they are restricted to a predefined set of concepts and more crucially, cannot ground them to image regions that these concepts represent. Recent methods [37, 58] have scaled CBMs to a large number of classes and also proposed to ways to learn multi-scale representations [52], but they are also unable to connect these concepts to image pixels. Our SSDG method addresses these challenges by producing local concept saliency maps, enhancing both DG performance and model prediction interpretability.

3. Methodology

The proposed framework is depicted in Figure 2. Training comprises three components: class prediction, concept prediction with saliency alignment, and localized contrastive training. During inference, class and concept predictions are integrated with a test-time correction mechanism.

In the concept alignment phase, we employ concept-level saliency maps as ground truth to enforce focus on interpretable features, using a saliency alignment loss. Domain invariance is further promoted by localized contrastive training, where we train the model to cluster similar concepts (e.g., `eyes` across augmentations) while separating unrelated concepts (e.g., `strings`). Finally, a test-time correction mechanism iteratively refines attention by suppressing irrelevant regions, leveraging concept-signatures to redirect focus and improve classification accuracy. We proceed with a concise review of the key notations, followed by an in-depth exposition of our approach and the pipeline for generating ground truth concept-level annotations.

We assume the source data is observed from a single domain, denoted as $D = \{x_i, y_i^c, y_i^k\}_{i=1}^N$, where x_i represents the i-th image, y_i^c the class label, y_i^k the concept label, and N the total sample count in the source domain. The shared CNN backbone is used to obtain the backbone features $F_x \in \mathbb{R}^{W \times H \times C}$. The automatically generated ground truth concept-level saliency maps are denoted by G_x^k. The GradCAM [43] saliency maps corresponding to class and concepts labels are denoted by S_x^c and S_x^k respectively.

3.1. Generating Concept-level Annotations

During training, we aim our model to identify and spatially attend to stable, discriminative regions. However, existing DG datasets lack fine-grained, concept-level annotations for such regions. To address this, we propose a novel pipeline that uses LLMs and DMs to automate scalable concept-level saliency map generation.

Our primary insight is that DMs can be harnessed to generate high-quality, concept-level saliency maps for synthesized images. Extracting cross-attention maps [1, 3] from a DM, given the prompt with concepts and corresponding synthesized image, yields highly granular concept-level saliency maps. Figure 3 demonstrates how DMs emphasize specific regions for distinct concepts, capturing fine-grained attention to features such as a cat's `whiskers` or a snowman's `hands`. With this technique serving as an efficient tool for yielding concept maps for synthesized images, the ensuing questions are: (i) how to automate the identification of pertinent concepts for each class across datasets, and (ii) how to transfer these concept maps to real-world images to generate ground-truth annotations.

First, to identify the key concepts associated with each class, we use GPT-3.5 [2], which we prompt to generate a

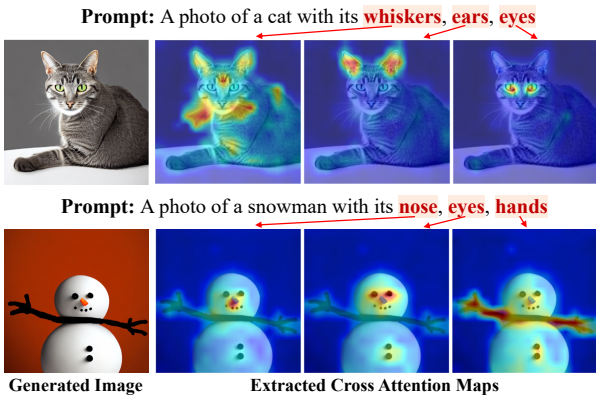

Figure 3. The first column displays the image generated from the given prompt, while the subsequent three columns show the cross-attention maps corresponding to each concept in the prompt.

list of distinctive, stable features (prompt in the supplement) for each class. For instance, GPT-3.5 outputs concepts such as `whiskers`, `ears`, and `eyes` for a cat. We generate a prompt leveraging these concepts, which is then used to synthesize an exemplar image for each class. We further derive the corresponding concept-level attention maps as outlined in the preceding paragraph.

Given a single synthesized exemplar image for each class and the concept-level saliency maps, we turn to the task of transferring these saliency annotations to real-world images. DM's feature space is particularly well-suited for this, as it captures detailed, semantically-rich representations that allow us to match synthetic concept regions to real images across domains. Leveraging this, we use the Diffusion Feature Transfer (DIFT) method [47], which computes pixel-level correspondences by comparing cosine similarities between DM features. Through this approach, we establish a region-to-region correspondence between synthetic saliency maps (e.g., `mouth` of a dog) and similar regions in real-world images. This enables us to generate comprehensive, consistent concept-level annotations across domains as shown in Figure 4.

We obtain these concept-level annotations for widely-used benchmarks, including VLCS [13], PACS [27], DomainNet [38], and Office-Home [50]. For an image x, the binarized concept-level maps (G_x^k) are henceforth referred to as ground-truth saliency maps (GT-maps) in this paper. Having established a process for generating saliency maps, the next challenge is identifying the subset of concepts that the model actually relies on for making predictions.

3.1.1 Concepts that matter

We restrict our method to essential concepts, filtering out those that do not contribute meaningfully to classification.

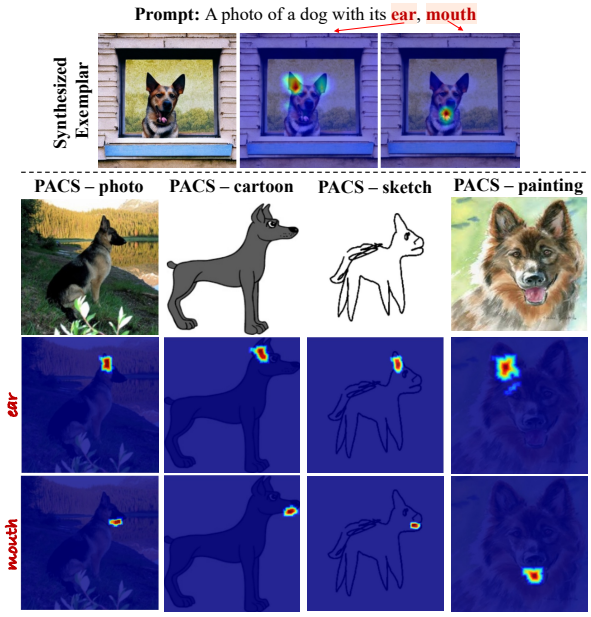

Figure 4. The first row presents the prompt, corresponding synthesized image and attention maps for the concepts `ear` and `mouth`. Below, we demonstrate that using diffusion features correspondences these concept saliency maps from a single exemplar can be automatically transferred on dog images across domains.

To do this, we train a ResNet-18 [18] classifier on the source domain [26], compute GradCAM [43] saliency maps for each class label, and use them to identify regions the model focuses on when making predictions.

We compute the overlap between the GradCAM maps and GT-maps for each known concept in the dataset. Given image x, its saliency map S_x^c for class c and the GT-map G_x^k for concept k, we define the overlap $O_c^k(x)$ as:

$$O_c^k(x) = \frac{\sum_{i,j} \min(S_x^c(i,j), G_x^k(i,j))}{\sum_{i,j} G_x^k(i,j)}. \tag{1}$$

where (i, j) are matrix indices. This measures how much of the model's attention for a given class aligns with the regions corresponding to concept k. We compute this overlap for all concepts and images in the training set, and for each class c, we define the set of important concepts $\mathcal{K}_c$ as those that consistently exhibit high overlap with the saliency maps:

$$\mathcal{K}_c = \left\{ k \;\middle|\; \frac{1}{N_c} \sum_{x \in \mathcal{D}_c} O_c^k(x) > \tau \right\}, \tag{2}$$

where N_c is the number of training examples in class c, $\mathcal{D}_c$ is the set of images in class c, and τ is a threshold that determines the importance of the concept. This procedure functions as a concept discovery module, identifying which

local concepts are critical for predicting each class.

3.2. TIDE

As shown in Figure 2, TIDE utilizes cross-entropy losses $\mathcal{L}_c$ and $\mathcal{L}_k$ for class and concept labels, respectively, integrated with novel concept saliency alignment and local concept contrastive losses. We detail these loss terms below.

3.2.1 Concept saliency alignment loss

For each image x, we predict important concepts $\mathcal{K}_c$ and enforce alignment between the saliency maps S_x^k for predicted concepts and the GT-maps G_x^k. This is encouraged by our proposed concept saliency alignment (**CSA**) loss:

$$\mathcal{L}_{\text{CSA}} = \frac{1}{|\mathcal{K}_c|} \sum_{k \in \mathcal{K}_c} \|S_x^k - G_x^k\|_2^2. \tag{3}$$

Aligning the model's attention with GT-maps enables class-specific reasoning, thereby elucidating the rationale behind predictions by linking them to relevant local features.

3.2.2 Local concept contrastive loss

While the CSA loss facilitates explicit localization of concepts, it is equally essential for these concept-level features to exhibit invariance across domain shifts. To achieve domain invariance, we propose a local concept contrastive (**LCC**) loss, employing a triplet strategy to cluster similar concepts (e.g., `eyes`) across domains while distinguishing unrelated ones (e.g., `feathers` and `ears`).

Let x be an anchor image containing concept k, a positive image x^+ (an augmentation of x that retains concept k), and a negative image x^- (containing a different concept k'). For each image, we compute a concept-specific feature vector $f_x^k \in \mathbb{R}^C$, emphasizing the concept's relevant regions using the GT-map G_x^k. Each element $f_x^k(l)$ is computed as:

$$f_x^k(l) = \sum_{i,j} G_x^k \cdot F_x(i,j,l). \tag{4}$$

where $\cdot$ is an element-wise multiplication operation, G_x^k is a matrix representing the GT-map, F_x represents convolutional feature maps from the last convolution layer, and (i,j,l) denotes the l^{th} channel's (i,j) element. These vectors focus on the concepts of interest (concept k for x and x^+, k' for x^-), in contrast to the global features used in prior works. The LCC loss is then defined as:

$$\mathcal{L}_{\text{LCC}} = \max(0, d(f_x^k, f_{x^+}^k) - d(f_x^k, f_{x^-}^{k'}) + \alpha), \tag{5}$$

where $d(.)$ is the euclidean distance and α is the margin.

Algorithm 1 Iterative Test-time Correction

Input: Test image x, initial class prediction $c^{(0)}$, signatures p^k for each concept k
Parameter: Threshold δ, max iterations T
Output: Corrected class prediction, c_{final}

$S_x^{c^{(0)},0} \leftarrow \text{GradCAM}(x, c^{(0)})$ {Initialize saliency map for iteration $t = 0$}
$x_{\text{masked}}^0 \leftarrow x$
for each $k \in \mathcal{K}_{c^{(0)}}$ **do**
  $f_x^{k,0} \leftarrow \sum_{i,j} S_x^k \cdot F_x(i,j,:)$
  $d(f_x^{k,0}, p^k) \leftarrow 1 - \frac{f_x^{k,0} \cdot p^k}{\|f_x^{k,0}\| \|p^k\|}$
  if $d(f_x^{k,0}, p^k) > \delta$ **then**
    Correction Phase:
    for $t \leftarrow 1$ **to** T **do**
      $M^{c^{(t-1)},t} \leftarrow \text{Binarize}(S_x^{c^{(t-1)},t-1})$
      $x_{\text{masked}}^t \leftarrow x_{\text{masked}}^{t-1} \cdot M^{c^{(t-1)},t}$
      $c^{(t)} \leftarrow \text{PredictClass}(x_{\text{masked}}^t)$
      $S_{x_{\text{masked}}}^{c^{(t)},t} \leftarrow \text{GradCAM}(x_{\text{masked}}^t, c^{(t)})$
      for each $k \in \mathcal{K}_{c^{(t)}}$ **do**
        $f_{x_{\text{masked}}}^{k,t} \leftarrow \sum_{i,j} S_{x_{\text{masked}}}^{k,t} \cdot F_x(i,j,:)$
        $d(f_{x_{\text{masked}}}^{k,t}, p^k) \leftarrow 1 - \frac{f_{x_{\text{masked}}}^{k,t} \cdot p^k}{\|f_{x_{\text{masked}}}^{k,t}\| \|p^k\|}$
        if $d(f_{x_{\text{masked}}}^{k,t}, p^k) \leq \delta$ **then**
          Return $c_{\text{final}} \leftarrow c^{(t)}$
        end if
      end for
    end for
  end if
  Return $c_{\text{final}} \leftarrow c^{(0)}$
end for

3.3. Test-time Correction

We establish that our localized, interpretable approach facilitates correction of misclassifications through concept-level feature verification. In this section, we first introduce concept signatures and detail the proposed correction strategy.

3.3.1 Local concept signatures

For each concept k, we define a concept-signature $p^k \in \mathbb{R}^C$, as its representative vector. We derive p^k by averaging the concept-specific feature vectors f_x^k across all training samples $x \in \mathcal{D}$ containing concept k (denoted as $\mathcal{D}^k$).

$$p^k = \frac{1}{|\mathcal{D}^k|} \sum_{x \in \mathcal{D}^k} f_x^k. \tag{6}$$

These vectors act as reference points, helping the model recognize when its attention is aligned with the right concepts during prediction, even when encountering new, unseen domains.

3.3.2 Detecting and Correcting misclassifications

Consider the example of test-time correction in Figure 2, where the model misclassifies a *person* as a *guitar*. The predicted class *guitar*, involves concepts like `strings` and

knobs, but these are absent in the image. As a result, the model erroneously focuses on irrelevant features (e.g., the person's legs or the background), as reflected in the Grad-CAM saliency maps.

To address misclassifications, we employ a two-step approach: possible mistake detection followed by correction. The process is outlined in Algorithm 1. First, the model extracts concept-level saliency maps (S_x^k), for all the concepts corresponding to the predicted class, which are then used to compute concept-level features f_x^k (step 4). In Step 5 and 6, we compare the features to the stored concept signatures; a deviation exceeding δ signals a misclassification.

Once a misalignment is detected, the model enters an iterative refinement phase. The features prominent for the current class level predictions are masked through the corresponding saliency map S_x^c (step 9 and 10). The masking process is cumulative, in effect utilizing all masks from the first to the t^{th} iteration. The masked features are then used for subsequent predictions (step 11). This process continues until the concepts corresponding to the predicted class aligns with concept-signatures or the maximum iteration count is reached (steps 8–16 in the algorithm). If alignment is achieved, the predicted class in that iteration is confirmed as the final output.

4. Experiments

Datasets: We conduct experiments on four widely used datasets - PACS [27], VLCS [13], OfficeHome [50], and DomainNet [38], within the DomainBed [17] evaluation benchmarks. PACS contains 9,991 images across 7 categories in four domains: 'sketch', 'photo', 'clipart', and 'painting'. VLCS consists of 10,729 images over 5 categories and 4 domains. The domains are from VOC2007 (V), LabelMe (L), Caltech101 (C), and SUN09 (S), with domain shifts primarily driven by background and viewpoint variations. OfficeHome, with 15,500 images across 65 categories, emphasizes indoor classes across 'product', 'real', 'clipart', and 'art' domains. For DomainNet, we follow prior work [39, 46] and use a subset of the 40 most common classes across 'sketch', 'real', 'clipart', and 'painting'.

Experimental Setup: We adhere to the SSDG paradigm, training on one *source* domain and testing across three *target* domains, i.e., the model is trained independently four times, with the averaged test accuracies across target domains reported in each case. Training utilizes solely source-domain GT-maps, excluding any target-domain prior knowledge, data or annotations. Minimal augmentations (quantization, blurring, and canny edge), were used to introduce slight perturbations to create the triplets. We use SDv2.1 [42] for generating exemplar images and computing cross-attention maps [1, 3]. To ensure a fair comparison, we adopt a ResNet-18 backbone throughout all the experiments. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a warm-up schedule over the first 1000 steps, after which the rate remains constant. Margin value α is set to 1.0. The batch size is set to 32. We empirically set $\delta = 0.1$ and cap test-time correction at $T = 10$ iterations.

Compared Approaches: We compare our method with ERM [17] baseline and existing approaches that utilize augmentation-based techniques (NJPP [59], AugMix [19], MixStyle [63], CutMix [60], RandAugment [7]), self-supervised and domain adaptation methods (RSC [22], pAdaIn [36], L2D [55], RSC+ASR [12]), uncertainty modeling (DSU [29], DSU-MAD [41]), attention and meta-learning methods (ACVC [8], P-RC [6], Meta-Casual [4]), and prompt-based learning (PromptD [28]). The number of methods evaluated differs by dataset. The discrepancy stems from PACS being the dominant benchmark in prior SSDG works, resulting in a larger number of methods evaluated on it. For VLCS, OfficeHome, and DomainNet, we rely on reported results from respective papers, or compute them ourselves (where code was available) to ensure a fair comparison.

4.1. Main Results

Table 1 provide a comparison of average classification accuracy for our approach against existing methods across the PACS, VLCS, OfficeHome, and DomainNet datasets. Each column in these tables represents a source domain used for training, with the numerical values indicating the average accuracy on the three target domains. The last column presents the average of the four columns.

TIDE decisively outperforms existing approaches across all datasets. It achieves average accuracy gains of 8.33%, 13.37%, 16.16%, and 8.84% over the second-best approach on PACS, VLCS, OfficeHome, and DomainNet, respectively. Table 1b details the performance on VLCS, where domain shifts primarily stem from background and viewpoint variations, with scenes spanning urban to rural and favoring non-standard viewpoints. Data augmentation based methods, which focus on style variation, yield limited gains on VLCS relative to PACS. Nonetheless, TIDE secures substantial improvements, notably achieving a 25.34% gain over DSU-MAD when Caltech101 (C) serves as the source domain. Another noteworthy observation is that TIDE maintains strong performance across varying class counts, from VLCS's 5 to OfficeHome's 65, underscoring the scalability of the proposed approach.

4.2. Ablations

Components of TIDE: Ablation experiments on the PACS dataset are conducted to assess the contribution of each loss

Method Venue	Art	Cartoon	Sketch	Photo	Average
ERM	65.38	64.20	34.15	33.65	49.35
Augmix [19] ICLR'21	66.54	70.16	52.48	38.30	57.12
RSC [22] ECCV'20	73.40	75.90	56.20	41.60	61.03
Mixstyle [63] ICLR'21	67.60	70.38	34.57	37.44	52.00
pAdaIn [36] CVPR'21	64.96	65.24	32.04	33.66	49.72
RSC+ASR [12] CVPR'21	76.70	79.30	61.60	54.60	68.30
L2D [55] ICCV'21	76.91	77.88	53.66	52.29	65.93
DSU [29] ICLR'22	71.54	74.51	47.75	42.10	58.73
ACVC [8] CVPR'22	73.68	77.39	55.30	48.05	63.10
DSU-MAD [41] CVPR'23	72.41	74.47	49.60	44.15	60.66
P-RC [6] CVPR'23	76.98	78.54	62.89	57.11	68.88
Meta-Casual [4] CVPR'23	77.13	80.14	62.55	59.60	69.86
ABA [5] ICCV'23	75.69	77.36	54.12	59.04	66.30
PromptD [28] CVPR'24	78.77	82.69	62.94	60.09	71.87
TIDE	**86.24**	**86.37**	**73.11**	**74.36**	**80.02**

(a) PACS

Method	V	L	C	S	Average
ERM	76.72	58.86	44.95	57.71	59.06
Augmix [19] ICLR'19	75.25	59.52	45.90	57.43	59.03
pAdaIn [36] CVPR'21	76.03	65.21	43.17	57.94	60.34
Mixstyle [63] ICLR'21	75.73	61.29	44.66	56.57	59.06
ACVC [8] CVPR'22	76.15	61.23	47.43	60.18	61.75
DSU [29] ICLR'22	76.93	69.20	46.54	58.36	62.11
DSU-MAD [41] CVPR'23	76.99	70.85	44.78	62.23	63.71
TIDE	**82.62**	**86.13**	**70.12**	**72.44**	**77.08**

(b) VLCS

Method	Art	Clipart	Product	Real	Average
ERM	57.43	50.83	48.9	58.68	53.96
MixUp [21] ICLR'18	50.41	43.19	41.24	51.89	46.93
CutMix [60] ICCV'19	49.17	46.15	41.2	53.64	47.04
Augmix [19] ICLR'19	56.86	54.12	52.02	60.12	56.03
RandAugment [7] CVPRW'20	58.07	55.32	52.02	60.82	56.56
CutOut [9] arXiv:1708	54.36	50.79	47.68	58.24	52.77
RSC [22] ECCV'20	53.51	48.98	47.16	58.3	52.73
MEADA [62] NIPS'20	57.0	53.2	48.81	59.21	54.80
PixMix [20] CVPR'22	53.77	52.68	48.91	58.68	53.51
L2D [55] ICCV'21	52.79	48.97	47.75	58.31	51.71
ACVC [8] CVPR'22	54.3	51.32	47.69	56.25	52.89
NJPP [59] ICML'24	60.72	54.95	52.47	61.26	57.85
TIDE	**72.32**	**75.13**	**75.22**	**73.37**	**74.01**

(c) OfficeHome

Method	Clipart	Painting	Real	Sketch	Average
ERM	68.73	66.12	68.51	69.44	68.2
MixUp [21] ICLR'18	70.31	64.34	69.21	68.82	68.17
CutMix [60] ICCV'19	71.52	63.84	67.13	69.41	67.98
Augmix [19] ICLR'19	72.37	62.91	69.84	71.22	69.09
RandAugment [7] CVPRW'20	69.71	65.51	68.36	66.93	67.63
CutOut [9] arXiv:1708	70.86	64.48	69.92	71.55	69.20
RSC [22] ECCV'20	68.25	67.91	70.76	66.18	68.28
PixMix [20] CVPR'22	72.12	63.51	71.34	65.46	68.12
NJPP [59] ICML'24	76.14	69.24	76.61	71.21	73.3
TIDE	**82.42**	**80.37**	**84.15**	**81.61**	**82.14**

(d) DomainNet

Table 1. SSDG classification accuracies (%) on PACS, VLCS, OfficeHome and DomainNet datasets, with ResNet-18 as backbone. Each column title indicates the source domain, and the numerical values represent the average performance in the target domains.

term on the overall performance. In Table 2, we present classification accuracy by incrementally adding components to the training pipeline, demonstrating the contribution of each element to the model's performance. We observe that the introduction of the concept classification loss $\mathcal{L}_k$ and concept saliency alignment loss $\mathcal{L}_{CSA}$ does not significantly affect the classifier's accuracy on the test domain. However, these components are integral to our approach, enabling test-time correction and, in turn, enhancing both performance and model interpretability. We observe that the introduction of the $\mathcal{L}_{LCC}$ loss leads to a substantial increase in test domain accuracy, clearly demonstrating its effectiveness in fostering domain invariance. Finally, we compare results before and after test-time correction, highlighting that even prior to correction, our method achieves state-of-the-art performance across all source domains on PACS, underscoring the strength of our approach.

Method	Art	Cartoon	Sketch	Photo	Average
$\mathcal{L}_c$	65.38	64.20	34.15	33.65	49.35
$\mathcal{L}_c + \mathcal{L}_k$	65.47	64.12	35.44	33.81	49.71
$\mathcal{L}_c + \mathcal{L}_k + \mathcal{L}_{CSA}$	66.11	64.23	35.14	34.27	49.93
$\mathcal{L}_c + \mathcal{L}_k + \mathcal{L}_{CSA} + \mathcal{L}_{LCC}$	80.28	82.91	66.12	65.37	73.67
+ test-time correction	**86.24**	**86.37**	**73.11**	**74.36**	**80.02**

Table 2. Ablation study (%) on PACS.

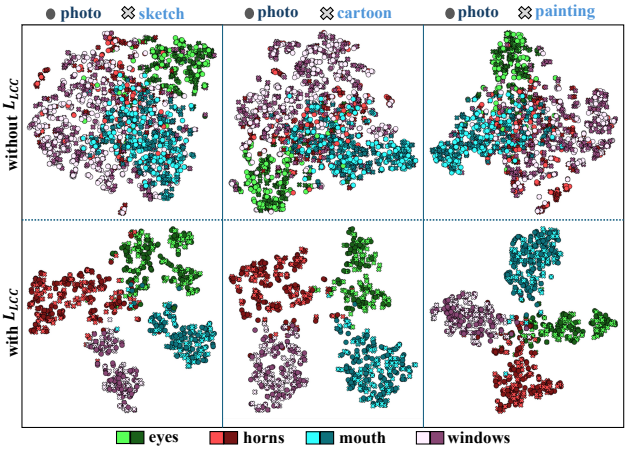

Figure 5. t-SNE visualizations to demonstrate impact of $\mathcal{L}_{LCC}$. Each column represents a test domain (Sketch, Cartoon, Painting), with the top row showing t-SNE plots without $\mathcal{L}_{LCC}$ applied and the bottom one with it. Please zoom in for optimal viewing.

Error analysis of test-time correction: For this analysis, we train TIDE on the photo domain of PACS and test it on the sketch domain. The model gives an initial accuracy of 74.79%, which improves to 82.29% post test-time correction. In 72.2% of test samples, TIDE does not invoke test-time correction, with class predictions correct in 93.8% of these cases, demonstrating substantial reliability. The supplement provides examples of cases where the model misclassifications are not picked-up in the signature matching step of test-time correction. In the remaining 27.8% of cases where correction is initiated, 52.5% successfully converge to the correct classification, significantly contributing

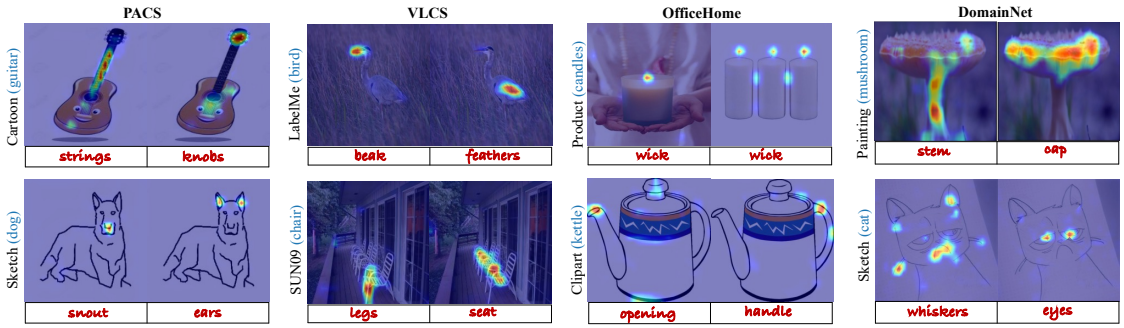

Figure 6. Illustrative examples of concept level GradCAM maps corresponding to TIDE's predictions, across the four studied datasets. The concept names are displayed beneath the maps, with the target domain and predicted class indicated on the left. More results in supplement.

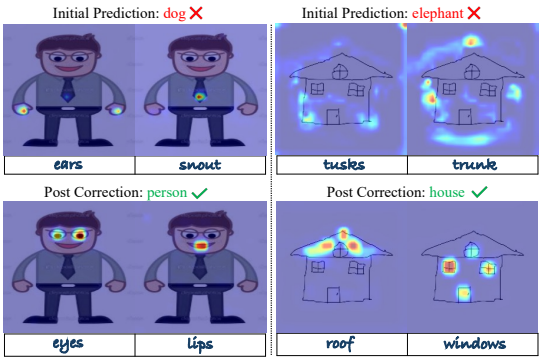

Figure 7. The top row shows initial class predictions and Grad-CAM maps for the concepts, while the bottom row presents the results after test-time correction.

to the final post-correction accuracy of 82.29%.

t-SNE plots: To demonstrate the impact of $\mathcal{L}_{LCC}$, we present t-SNE visualizations in Figure 5, showing the distribution of concept-specific vectors (as computed in Equation 4). We individually plot the source domain (photo) along with sketch, cartoon, and painting domains as targets. Each case includes two plots: one without $\mathcal{L}_{LCC}$ (top row) and one with it (bottom row). We observe that with $\mathcal{L}_{LCC}$, concept samples (e.g., `mouth`) align closely across domains, while distinct separation occurs between different concepts (e.g., `mouth` and `horns`). Without $\mathcal{L}_{LCC}$, cluster separation and alignment across domains are weaker, highlighting its role in improving intra-concept compactness and inter-concept separability.

4.3. Qualitative Results

Concept Localization: We train the SSDG model on photo domain for PACS, OfficeHome, and DomainNet and on Caltech101 for VLCS. The predicted class, concepts and corresponding saliency maps are shown in Figure 6. For each target class (e.g., mushroom, kettle, guitar), the model reliably highlights key concept-specific regions (e.g., `stem`, `handle`, `strings`) essential for classification. The model effectively isolates key features across diverse contexts, such as the `legs` and `seat` of a chair in complex zoomed-out scenes, the `beak` and `feathers` of camouflaged birds, the `eyes` and `whiskers` in deformed sketches of cats, and varying feature sizes, with the `wick` occupying a small area and the `mushroom cap` spanning a larger one.

Test-time Correction: The Figure 7 visually illustrates the efficacy of TIDE's test time correction abilities, by comparing the initial and corrected results. The model initially misclassifies the images as dog and elephant, resulting in poorly aligned concept-level saliency maps for corresponding concepts i.e. `ears`, `snout`, `tusks` and `trunk`. TIDE detects and rectifies such errors during inference, leading to accurate classification accompanied by precise concept localization. This qualitative evaluation reinforces the robustness of our approach in refining predictions and generating reliable, class-specific concept maps.

5. Conclusion

In this work, we considered the problem of single-source domain generalization and observed that the current state-of-the-art methods fail in cases of semantic domain shifts primarily due to the global nature of their learned features. To alleviate this issue, we proposed TIDE, a new approach that not only learns local concept representations but also produces localization maps for these concepts. With these maps, we showed we could visually interpret model decisions while also enabling correction of these decisions at test time using our iterative attention refinement strategy. Extensive experimentation on standard benchmarks demonstrated substantial and consistent performance gains over the current state-of-the-art. Future work will rigorously explore methods to generate confidence scores grounded in TIDE's concept verification strategy and would explore the application of TIDE on other (e.g. fine-grained) classification datasets.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 6

[2] Tom B Brown. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[3] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3, 6

[4] Jin Chen, Zhi Gao, Xinxiao Wu, and Jiebo Luo. Meta-causal learning for single domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 7

[5] Sheng Cheng, Tejas Gokhale, and Yezhou Yang. Adversarial bayesian augmentation for single-source domain generalization. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 7

[6] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 1, 2, 6, 7

[8] Ilke Cugu, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6, 7

[9] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552, 2017. 7

[10] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[11] Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. In *The German Conference on Pattern Recognition*, 2019. 2

[12] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6, 7

[13] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 4, 6

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[16] Tejas Gokhale, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. Attribute-guided adversarial training for robustness to natural perturbations. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021. 2

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *International Conference on Learning Representations (ICLR)*, 2020. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[19] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 6, 7

[20] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 7

[21] Yann N. Dauphin David Lopez-Paz Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations (ICLR)*, 2018. 7

[22] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6, 7

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 2

[24] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020. 3

[25] Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(3):766–785, 2019. 1

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 4

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *International Conference on Computer Vision (ICCV)*, 2017. 4, 6

[28] Deng Li, Aming Wu, Yaowei Wang, and Yahong Han. Prompt-driven dynamic object-centric learning for single domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6, 7

[29] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *International Conference on Learning Representations (ICLR)*, 2022. 2, 6, 7

[30] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[31] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *ArXiv*, abs/2108.13624, 2021. 1

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015. 1

[33] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *International Conference on Learning Representations (ICLR)*, 2021. 3

[34] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[35] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, 2013. 2

[36] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6, 7

[37] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *International Conference on Learning Representations (ICLR)*, 2023. 3

[38] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision (ICCV)*, 2019. 4, 6

[39] Daiqing Qi, Handong Zhao, Aidong Zhang, and Sheng Li. Generalizing to unseen domains via text-guided augmentation. In *European Conference on Computer Vision (ECCV)*, 2024. 6

[40] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[41] Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, and Tao Mei. Modality-agnostic debiasing for single domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6, 7

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3, 4

[44] Sarath Sivaprasad, Akshay Goindani, Mario Fritz, and Vineet Gandhi. Class-wise domain generalization: A novel framework for evaluating distributional shift. In *NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 1, 2

[45] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *ArXiv*, abs/2006.11207, 2020. 2

[46] Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: An empirical odyssey. In *European Conference on Computer Vision Workshops (ECCVW)*, 2020. 6

[47] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 4

[48] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[49] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 6

[51] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1

[52] Bor-Shiun Wang, Chien-Yi Wang, and Wei-Chen Chiu. Mcpnet: An interpretable classifier via multi-level concept prototypes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[53] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *International Conference on Learning Representations (ICLR)*, 2019. 2

[54] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35 (8):8052–8072, 2022. 1

[55] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *International Conference on Computer Vision (ICCV)*, 2021. 6, 7

[56] Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees Snoek. A bit more bayesian: Domain-invariant learning with uncertainty. In *International Conference on Machine Learning (ICML)*, 2021. 2

[57] Qinwei Xu, Ruipeng Zhang, Yi-Yan Wu, Ya Zhang, Ning Liu, and Yanfeng Wang. Simde: A simple domain expansion approach for single-source domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[58] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[59] Jianhao Yuan, Francesco Pinto, Adam Davies, and Philip Torr. Not just pretty pictures: Toward interventional data augmentation using text-to-image generators. *International Conference on Machine Learning (ICML)*, 2024. 1, 2, 6, 7

[60] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 6, 7

[61] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[62] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7

[63] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 6, 7

[64] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision (IJCV)*, 132(3):822–836, 2024. 2