Domain Gap Embeddings for Generative Dataset Augmentation

Yinong Oliver Wang* Younjoon Chung* Chen Henry Wu Fernando De la Torre
Carnegie Mellon University
{yinongwa, younjooc, chenwu2, ftorre}@cs.cmu.edu

Abstract

The performance of deep learning models is intrinsically tied to the quality, volume, and relevance of their training data. Gathering ample data for production scenarios often demands significant time and resources. Among various strategies, data augmentation circumvents exhaustive data collection by generating new data points from existing ones. However, traditional augmentation techniques can be less effective amidst a shift in training and testing distributions.

*This paper explores the potential of synthetic data by leveraging large pre-trained models for data augmentation, especially when confronted with distribution shifts. Although recent advancements in generative models have enabled several prior works in cross-distribution data generation, they require model fine-tuning and a complex setup. To bypass these shortcomings, we introduce **Do**main **G**ap **E**mbeddings (DoGE), a plug-and-play semantic data augmentation framework in a **cross-distribution few-shot** setting. Our method extracts disparities between source and desired data distributions in a latent form, and subsequently steers a generative process to supplement the training set with endless diverse synthetic samples. Our evaluations, conducted on a subpopulation shift and three domain adaptation scenarios under a few-shot paradigm, reveal that our versatile method improves performance across tasks without needing hands-on intervention or intricate fine-tuning. DoGE paves the way to effortlessly generate realistic, controllable synthetic datasets following the test distributions, bolstering real-world efficacy for downstream task models.*

1. Introduction

The swift progression of computer vision in the past decade can be attributed to improved deep learning algorithms for large-scale training, increased computing power, and the availability of vast datasets such as ImageNet [15] and LAION-5B [66]. While such internet-scale real-world datasets allow to train general vision models, they are not tailored to application scenarios with specific data distributions, *i.e.*, the **cross-distribution** adaptation, which can lead to serious concerns in reliability [41, 63]. This issue often requires costly data collection where models operate.

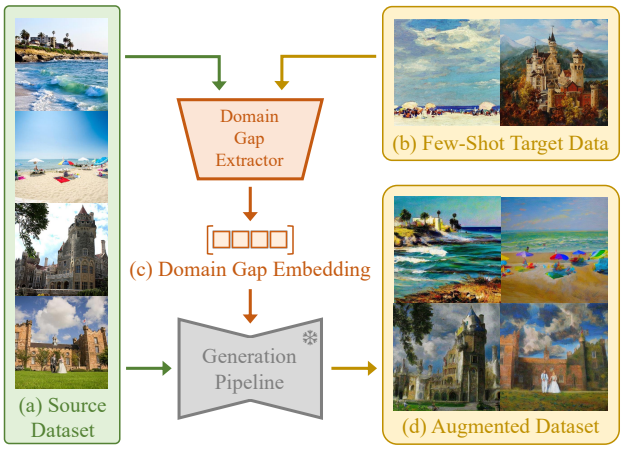

Figure 1. **Overview**. In real-world applications, computer vision models often suffer from discrepancies between training and testing data distributions. To alleviate this problem, we propose a novel dataset augmentation method to complement the training dataset with synthetic images. Given (a) a source dataset (*e.g.*, real photos), and (b) a few samples from a target distribution (*e.g.*, paintings), we extract the distribution differences into (c) Domain Gap Embeddings, which enables generating (d) augmented synthetic data to enhance the model performance.

Among various solutions to this issue, data augmentation has been explored to alleviate such extensive data collection. However, images generated with traditional data augmentation through flipping, gamma adjustments, noise, or more sophisticated methods [14, 81] often fail to align the augmented data with shifted test distributions. Although there are cross-domain augmentation techniques [43, 44], these strategies are task-specific and not easily transferable to other problems. Besides these data-centric efforts, unsupervised domain adaptation (UDA) is an active field of research for such problems from the model aspect (*e.g.*, [32]). Our approach distinguishes itself from the above by its ability to produce *endless* data with much more variability and the need for much fewer samples (*i.e.*, **few-shot**).

To mitigate the distribution discrepancy issues, synthetic

*Equal contribution.
†Source code available at https://domain-gap-embeddings.github.io/.

datasets have also been studied as a more controllable, diverse, high-quality supplement to the training dataset. Traditionally, simulators and graphics engines are the primary sources of synthetic datasets [56, 70, 76]. However, they typically suffer from unrealism (*i.e.*, domain gap) and bounded diversity [25]. With the advancement of visual generative models, they are leveraged for in-domain dataset synthesis in recent works [1, 34, 83]. Nonetheless, very few dataset generation methods [3] focus on the cross-distribution setting guided by just a few input target samples (*e.g.*, 20 images), which is realistic in many scenarios of interest. Moreover, to the best of our knowledge, none achieves target dataset synthesis in such a setting without fine-tuning. The question that we try to address in this paper is: *Can we use off-the-shelf large pre-trained models (LPMs) as synthetic data generators for effective few-shot dataset augmentation towards specific data distributions?*

To address this question, we propose DoGE, a few-shot cross-distribution dataset generation framework that is task-agnostic and **inference-only**, as shown in Figure 1. The framework takes (a) a source distribution (*i.e.*, the original training dataset), and (b) a few samples from a target distribution in the application context. We propose to extract the distribution discrepancies (*e.g.*, semantic changes, style transfer) into (c) representations in the CLIP latent space [58], named the Domain Gap Embeddings. We then utilize the extracted gap representations to augment source data to generate (d) synthetic datasets that follow the same distribution as the provided few target images.

Our method successfully generates synthetic supplementary datasets as long as (1) the latent representation space, CLIP, has the capacity to express the distribution differences, and (2) the generative diffusion models, Stable UnCLIP [61], is capable of generating in the target distribution. Under these loose constraints, we show that our synthetic datasets from DoGE significantly improve model performance in various computer vision tasks, including subpopulation shifts and domain adaptation. Moreover, DoGE is compatible with and complementary to parallel methods such as UDA and fine-tuning. In summary, DoGE provides the following contributions:

- **Accessibility**: Our framework offers a plug-and-play dataset augmentation experience. With a source dataset to augment, users only need to provide *a few unlabeled images from the target distribution* to obtain an effective synthetic dataset in the desired domain.
- **Efficiency**: Our cross-distribution dataset augmentation framework generates data in the target domain *without the need for fine-tuning*. We directly take advantage of public LPMs, and each step can be inference-only.
- **Effectiveness**: The synthetic datasets from our generation pipeline can successfully improve the task model performance by a significant margin.

2. Related Works

While real-world images are cornerstones of computer vision, as modern vision datasets increase in size, it has become gradually more challenging to scrutinize and clean the collected data. The difficulty of curating large datasets poses potential issues such as noisy labels and dataset imbalance [5, 8, 52]. Hence synthetic data became a popular alternative with high controllability and accessiblility.

Generative Models for Image Data: Recent advances in generative models have provided powerful tools for synthetic data generation. Generative Adversarial Network (GAN) pioneered a new direction for high-quality image synthesis [6, 21, 37, 38]. In parallel, diffusion models [28, 29, 51, 69] demonstrate their promising potential, leading to many astonishing works including GLIDE [50], DALL•E 2 [59], Imagen [64], and Stable Diffusion [61].

Besides generative backbones, fine-grained controllability of the generative models is also essential for data synthesis. In the direction of GANs, CycleGAN [86], CyCADA [30], and CLIP-enabled methods [54, 87] achieved effective image-to-image transfer and targeted editing toward desired distributions. For diffusion models, various conditioning techniques regulate the generations. Some methods [7, 26] leverage the cross-attention maps to apply accurate prompt-based augmentations. Other works [18, 22, 75, 80] learn special tokens and embeddings to preserve identities during data generation. Similarly, methods in [23, 33, 39, 40, 62] fine-tune the diffusion models for desired generation, while image-to-image synthesis [49, 77] is also critical to data augmentation. Finally, ControlNet [82] uses condition maps to control the generation accurately.

Synthetic Data Generation: With such extensive generation capability and fine-grained controllability, generative models have been leveraged to populate synthetic datasets [9, 34]. GANs have been used for effective synthetic dataset generation through latent space manipulation [4, 42, 47, 83]. Enabled by the abundant generation controls in diffusion-based networks, more recent works leverage diffusion models to improve data diversity by expanding existing datasets [2, 57, 68, 71, 74, 84].

While the above methods can generally expand a given dataset, they suffer from subpopulation and domain shifts in datasets. Regarding subpopulation shifts, Fill-Up [67] incorporates Textual Inversion [18] to fix imbalanced datasets by uneven generation but requires optimizing a token for each class. To address domain shifts, some methods [16, 79] utilize captioning models to describe target distributions and construct new prompts for generation. However, since the expressibility of texts is limited, other methods also resort to fine-tuning for adaptation. Assuming access to the full target dataset, solutions in [1, 55] fine-tuned Imagen and DDPM [29] for better in-domain generations. Under the

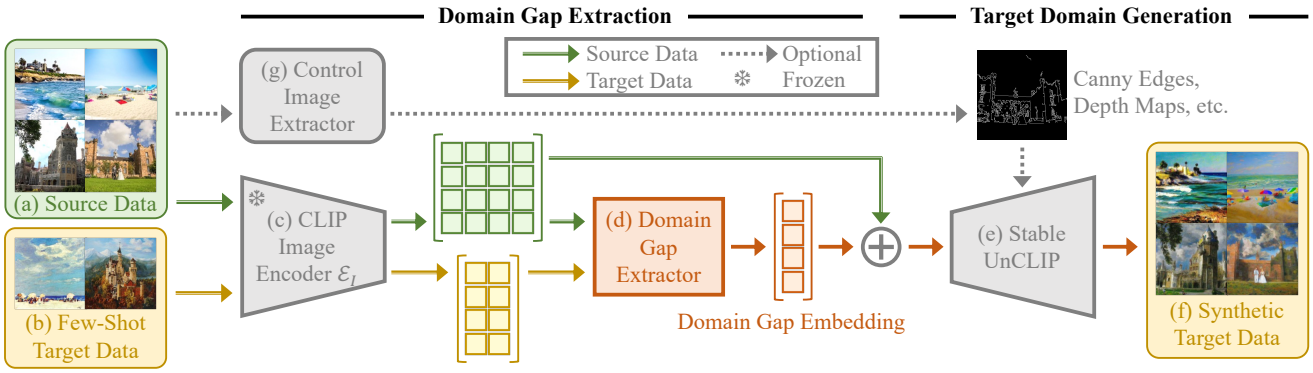

Figure 2. **Framework**: (a) The source dataset and (b) a few target data samples are first (c) encoded in the CLIP embedding space. We then (d) extract the representation, named the Domain Gap Embedding, between the source and target distributions. The Domain Gap Embedding augments source image embeddings to construct the latent input to (e) the generative model (Stable UnCLIP), which generates (f) a synthetic dataset following the target distribution. Optionally (dotted lines), we can (g) integrate ControlNet to provide further structural guidance to preserve the source image structures.

few-shot setup where only a few target samples are available, DomainStudio [85] introduced similarity loss to conquer the over-fitting issue in fine-tuning, and DATUM [3] proposed to fine-tune the model into the target domain with crops of the few target samples. Nonetheless, such methods require domain-specific fine-tuning and may introduce training algorithm modification, while our method, with better performance in our experiment setups, can be directly applied off the shelf for the given target images.

3. Method

Recognizing the lack of practical and readily available cross-distribution dataset synthesis methods, we introduce a novel, model-agnostic few-shot dataset augmentation framework. Our framework possesses the ability to create synthetic samples that conform to the target distribution based on a minimal set of input images. It is characterized by its simplicity and effectiveness, and, in its fundamental configuration, does not necessitate any training.

Our framework consists of two main components: modeling the domain gap and generating across the domain gap, shown in Fig. 2. To generate from one dataset distribution to another, we first capture the differences between them as Domain Gap Embeddings, shown in Sec. 3.1. With the representation for the distribution gap, Sec. 3.2 illustrates our method for generating datasets from the source to target distribution, with an optional trick to preserve image quality. To further improve the usefulness of the generated dataset, we also conduct confidence-based generation cleaning methods on downstream tasks, shown in Sec. 3.3.

3.1. Domain Gap Extraction

When capturing differences in data distributions, fine-tuning generative models across domains can be costly, and prompts may not articulate the discrepancies. Hence, we focus on modeling the distribution differences in the latent space. The recent research in visual representation learning introduces powerful semantic latent spaces such as CLIP. CLIP is assumed to have sufficient knowledge generalization for common settings, and its linear vector compositionality enables semantically meaningful operations [72]. In our framework, we choose to leverage the CLIP latent space to capture the gap between data distributions and directly apply it in data augmentation. Such captured distribution discrepancies are named the Domain Gap Embeddings.

Fig. 2 (left) shows the domain gap extraction process. The input consists of a source dataset $\mathcal{D}_\mathcal{S}$ (Fig. 2a), with $|\mathcal{D}_\mathcal{S}| = N$, and a few data samples $\mathcal{D}_\mathcal{T} = \{y_j\}_{j=1}^m$ (Fig. 2b) from a different target distribution with $m \ll N$. We first encode images from a randomly sampled subset $\hat{\mathcal{D}}_\mathcal{S} = \{x_i\}_{i=1}^n \subseteq \mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ into the CLIP space via a CLIP image encoder $\mathcal{E}_\mathcal{I}$ (Fig. 2c). Denoting the image embeddings as $z_{x_i} = \mathcal{E}_\mathcal{I}(x_i)$ and $z_{y_j} = \mathcal{E}_\mathcal{I}(y_j)$, we study two options as the Domain Gap Extractor (Fig. 2d) to capture the gap representation Δz. A straightforward way is computing the expected differences of all pairs between the source and target dataset, which is equivalent to the difference of the means of the images assuming $\mathcal{D}_\mathcal{S}$ is independent of $\mathcal{D}_\mathcal{T}$, *i.e.*,

$$\Delta z = \mathbb{E}_{x_i \in \mathcal{D}_\mathcal{S}}[\mathbb{E}_{y_j \in \mathcal{D}_\mathcal{T}}[\mathcal{E}_\mathcal{I}(y_j) - \mathcal{E}_\mathcal{I}(x_i)]] \quad (1)$$

$$= \frac{\sum_{j=1}^m z_{y_j}}{m} - \frac{\sum_{i=1}^n z_{x_i}}{n}. \quad (2)$$

Another way to extract the gap is through Principal Component Analysis (PCA) [17]. Since the first principal direction from PCA denotes the direction where a distribution varies the most, we leverage this property and apply PCA on a joint set $\{z_{x_i}\}_{i=1}^n + \{z_{y_j}\}_{j=1}^m$ with $n = m$. The first principal direction from PCA is then considered as the domain gap representation Δz. From empirical results shown in Appendix A, we observe that the first option of computing the domain gap (Eq. (2)) yields better generation quality. The impact of the values of n and m is also addressed in

Appendix A. For the rest of this paper, we adopt this mechanism as our Domain Gap Extractor, yet users can easily design and swap in their own extractor in our framework.

3.2. Target Dataset Generation

With the domain gap extracted into the latent form Δz, we augment source images to generate the synthetic dataset, as shown in the right half of Fig. 2. Capturing the gap Δz between distributions in the CLIP space opens up methods to generate data across domains. While various diffusion-based generative models accept texts or images as input or conditions, our method contrasts with these in that we directly interact with CLIP latent embeddings. Such is made possible by the UnCLIP approach introduced in DALL•E 2, specifically the Stable UnCLIP model [61] (Fig. 2e). It is a fine-tuned Stable Diffusion model that accepts CLIP image embeddings directly as input. Hence, in this paper, we choose the Stable UnCLIP model, denoted as G as our generation backbone in the framework.

Given the distribution gap Δz and source image embeddings $\{z_{x_i}\}$ both in the CLIP space, we augment the source image representations by simply adding the gap Δz to them. To further increase diversity, we also introduce small Gaussian random perturbations $\epsilon \sim \mathcal{N}(\mathbf{0}, 10^{-3}I)$ in the augmentation. Similarly, we introduce a distributional edit strength scalar $C \sim \mathcal{N}(c, 0.05)$. The impact of values of c is discussed in Sec. 4.4.2. Hence, the generated k images, denoted as $\{\hat{y}_i\}_{i=1}^{k}$ (Fig. 2f) are obtained as:

$$\hat{y}_i = G(z_{x_i} + C \cdot \Delta z + \epsilon). \tag{3}$$

The above two steps form our base framework and can already achieve effective target dataset generation (details in Sec. 4). Nonetheless, in specific cases, additional techniques can be adopted for higher generation quality.

Finer Generation Control: In some cases, it is beneficial to preserve the visual structure of the original source data to be augmented. For example, maintaining the object structure can further ensure less deformation or corruption in the generation. Because the expressiveness of one vector in the CLIP space is limited, a fine-grained structural control can introduce more detailed visual guidance on top of our domain gap embeddings.

Therefore, we integrate ControlNet into our generative module for accurate image structure control during the generation. As shown in Fig. 2g, we can feed the input source data through a control image extractor. Given a source image, this module outputs a series of domain-invariant control maps for generation, including Canny edge maps [10] and HED edge maps [78], and processing depth and segmentation maps from ground truth labels of source data, if available. During the generation phase, we feed these control maps into our revised Stable UnCLIP model for more refined generations. This guidance enriches our augmentation with the compositional information, which empirically brings further improvements as shown in Sec. 4.3.

3.3. Confidence-Based Generation Cleaning

While the ControlNet integration preserves the structure and quality of our generated datasets, it is not safe to assume that every synthetic image is valid and helpful data. Inspired by [53], we propose a confidence-based filtering mechanism to remove such poor generations. Given a downstream task model trained on the source data, *e.g.*, a classifier, we fine-tune this model with our data augmentation to improve test performance. At each iteration during fine-tuning, before training we first perform inference with the current model to filter out augmented data with highly confident but incorrect predictions. The confidence is determined as the highest predicted softmax score s among all classes. With a threshold parameter t, we discard synthetic samples where the model prediction is wrong but its confidence is greater than the threshold, *i.e.*, $s > t$. We only temporarily discard samples in each training step but never eliminate any data from the dataset. Please see Appendix B for details.

4. Experiments

This section illustrates the versatility and efficacy of DoGE, demonstrating its ability to produce synthetic datasets benefiting various computer vision challenges. Sec. 4.1 presents the standard experimental setups employed in our studies. Subsequently, Sec. 4.2 addresses issues related to imbalanced class distributions and showcases effectiveness under the presence of spurious correlations. In Sec. 4.3, we delve into the effectiveness of our dataset generation approach under common domain adaptation problems. In addition to task-based evaluations, we conduct ablation studies concerning our generative pipeline in Sec. 4.4. These studies serve to provide both qualitative and quantitative assessments of the synthetic datasets created through DoGE.

4.1. Experimental Setup

Baselines: For classification tasks in Sec. 4.2 and Sec. 4.3, *base* refers to the models trained on the source data in the cross-domain setting only. Subsequently, we fine-tuned the *base* models on the augmented datasets to assess the efficacy of data generation methods. We compared against one traditional augmentation, RandAugment [14], and two generative methods, DA-Fusion [71] and DATUM [3]. For fair comparisons, we kept the number of generated images the same within each task across all generative methods.

Implementation: For *base* classification models we fine-tuned ImageNet pre-trained ResNet50 [24] models for 20 epochs with AdamW optimizer [46] at a constant learning rate of 10^{-3} and a batch size of 128. For each generative

Figure 3. **Examples of synthetic CelebA data generated from (a) Source into (b) Target distribution.** Under subpopulation shift, we generated data from the majority subpopulation (a) into the under-represented distribution (b). (c) shows the synthetic data generated from our pipeline. The results demonstrate our capability to apply semantic augmentation in accordance with gaps between two distributions.

Method	Test Accuracy (%)
Base	38.00
Oversampling	53.80
RandAugment [14]	62.40
DA-Fusion [71]	59.72
DoGE (Ours)	**67.16**

Table 1. **Test Accuracy on our constructed CelebA imbalanced classification problem.** We evaluated our method against four baselines. This table shows that synthetic data from DoGE has a significant advantage over other methods.

Method	Test Accuracy (%)
Base	38.00
LoRA [33]	56.05
LoRA + DoGE (Ours)	**74.28**

Table 2. **Test Accuracy on CelebA imbalanced classification problem with fine-tuned generative models.** We applied our method on top of a personalized generator via LoRA and show that DoGE is complementary to adaptation via personalization.

baseline and our method, the *base* model is further fine-tuned on respective augmented dataset for 20 epochs, with AdamW optimizer and a batch size of 256. For CelebA, we used a constant learning rate of 10^{-3}. For DomainNet and FMoW, the classification head was trained with a learning rate of 10^{-4}, and its preceding layers with 10^{-5}. For all datasets, confidence-based generation cleaning was applied at training time with a threshold $t = 0.9$. After fine-tuning, the final models were saved for evaluation. We also extended to segmentation problems where we directly adopted the synthetic data evaluation pipeline generously published in DATUM, the current state-of-the-art method in one-shot UDA for self-driving segmentation problems.

4.2. Subpopulation Shift

In our initial experiment, we sought to assess the effectiveness of our solution in addressing the subpopulation shift problem. Specifically, we aimed to evaluate how well DoGE could mitigate imbalanced training data distributions that result in spurious correlations.

We curated subsets of facial data from the CelebA dataset, intentionally introducing imbalances in certain attributes. Given an attribute (*e.g.*, perceived gender), we selected a secondary attribute (*e.g.*, eyeglasses), as the bias factor. Then, we sampled 1000 males wearing eyeglasses and 1000 females without eyeglasses, denoted as the source (majority distribution) in Fig. 3a. We also sampled 20 images per class with the opposite secondary attribute (*i.e.*, bias), denoted as the target (minority distribution) in Fig. 3b. Training a perceived gender classification model on this imbalanced subset naturally introduced bias toward eyeglasses over gender.

To supplement this imbalanced training set, we first extracted the distribution gap for each class from randomly sampled only 10 source and 10 target images. Then, DoGE generated 1000 synthetic samples per class in the target (minority) distributions with $c = 1.0$. Fig. 3c shows the generated images following the target distribution where eyeglasses are successfully added or removed respectively to follow the under-represented subpopulation.

After the data generation, our new training set consists of 1000 sampled source data (Fig. 3a) and 1000 generated target data (Fig. 3c). For the test set, we sampled 1000 images per class from the target (minority) distribution (Fig. 3b) in CelebA. For comparison, we first oversampled target data by duplication, then applied RandAugment to this oversampled training set. DA-Fusion was also used to expand each class in the target data. All baselines generated the same amount of data in the evaluation as ours. Tab. 1 shows the test accuracy after training on our synthetic data along with the baseline performances to compare with. DoGE achieved the best test accuracy among the baselines.

Since fine-tuning is studied as a powerful method for targeted generation, we demonstrated our compatibility with LoRA [33] and generated synthetic data using a fine-tuned Stable UnCLIP model. The results in Tab. 2 indicate that our method complements adaptation via fine-tuning.

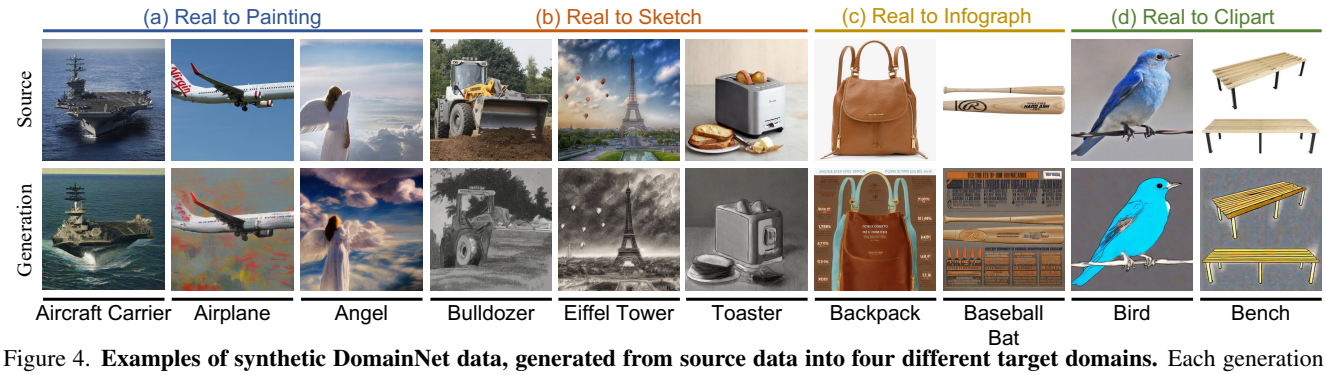

Figure 4. **Examples of synthetic DomainNet data, generated from source data into four different target domains.** Each generation (bottom) was augmented from the source image (top) using our pipeline with ControlNet. The results demonstrate our capability to augment data in accordance with gaps between distributions.

Method	DomainNet Acc (%)				FMoW Acc (%)		
	Painting	Infograph	Clipart	Sketch	Asia	Americas	Oceania
Base	34.64	14.48	39.06	24.70	66.27	64.65	74.42
RandAugment [14]	37.20	15.90	41.08	26.26	64.35	70.47	74.29
DA-Fusion [71]	39.57	16.54	42.22	28.27	70.97	76.71	77.32
DATUM [3]	38.19	17.80	40.96	29.46	71.63	78.35	75.24
DoGE (Ours)	**44.00**$^{\pm.0}$	**18.71**$^{\pm.0}$	**45.61**$^{\pm.3}$	**34.96**$^{\pm.3}$	**72.62**$^{\pm.1}$	**78.94**$^{\pm.2}$	**78.14**$^{\pm.1}$

Table 3. **Test accuracy in unsupervised domain adaptation classification problems.** We evaluated against four baselines on the left column. For DomainNet, the task is to adopt a model with a Real domain training dataset to Painting, Inforgraph, Clipart, and Sketch domains. For FMoW, for each region (Asia, Americas, Oceania), we adopted a model with old satellite images (2002-12) to perform well on new satellite data (2016-17). The table shows that our methods achieved the highest test accuracy in every category.

Method	Test Acc (%)	Δ
Base	32.86	—
DoGE	38.64	+5.78
DoGE + ControlNet	40.29	+1.65
DoGE + ControlNet + Cleaning	41.30	+1.01

Table 4. **Incremental improvements on DomainNet (Real → Painting) problem.** We gradually added our components to the base model and evaluated the effectiveness of each part.

UDA Method	Test Acc (%)		
	w/o DoGE	w/ DoGE	Δ
BSP [11]	46.76	47.34	+0.58
DANN [19]	47.01	49.68	+2.67
CDAN [45]	51.66	52.11	+0.45
MCD [65]	50.88	52.14	+1.26
MCC [35]	50.08	52.95	+2.87
MemSAC [36]	52.27	54.16	+1.89

Table 5. **Test Accuracy of UDA methods on the DomainNet (Real → Painting) problem.** We evaluated existing UDA methods with and without DoGE. The table shows that our approach is compatible with and complementary to UDA methods.

4.3. Unsupervised Domain Adaptation

4.3.1 Classification Tasks

DomainNet consists of 0.6 million images of 345 classes distributed across 6 unique domains including Real (**R**), Clipart (**C**), Infograph (**I**), Painting (**P**), Quickdraw (**Q**) and Sketch (**S**). We evaluated our method on 4 domain adaptation tasks from **R** to **P**, **S**, **C** and **I**, using official test sets. In each task, we randomly sampled 345 images from both the source (*i.e.*, **R**) and target (*i.e.*, **P**, **S**, **C** or **I**) distribution to calculate the domain gap. Synthetic **P** and **S** images were generated with edit strength mean $c = 1.3$, **I** with $c = 1.1$, and **C** with $c = 1.5$. Sec. 4.4.2 discusses the choice of values for c. For each class DoGE generated 128 images with ControlNet (Fig. 4) to supplement the training data for fine-tuning, increasing the dataset size by approximately 30%.

Tab. 4 shows the incremental improvements of training on Real domain and testing on Painting domain. Standalone DoGE improved w.r.t standard approaches and additional techniques further increased our advantages. Tab. 3 shows full comparisons on all four domains, and DoGE achieved the best accuracy in all settings.

To demonstrate DoGE's compatibility and improvements to traditional UDA solutions, we evaluated our performance based on six UDA methods. For each method, we incorporated DoGE by simply adding our synthetic images to the training dataset. Tab. 5 shows that our method can help further improve UDA methods in general. Please see Appendix D for the complete experiment.

Figure 5. **Examples of synthetic self-driving data generated from (a) GTA5 source images into (b) Cityscapes target domain.** (c) shows the synthetic data generated from our pipeline without any improvement tricks. We also demonstrated the generation with scene structure preserved by ControlNet (conditioned on canny edges and source segmentation ground truth) in (d). The synthetic data are then used in unsupervised domain adaptation methods to adapt models across domains.

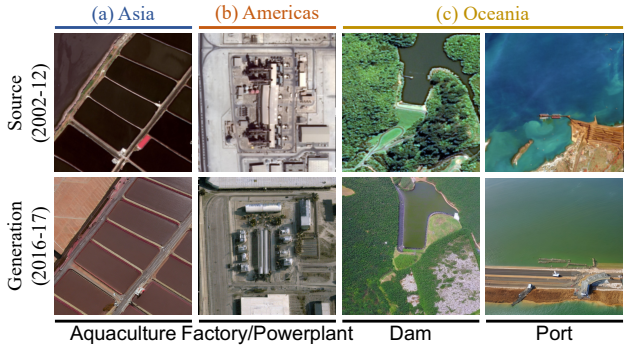

Figure 6. **Examples of synthetic FMoW data**, generated from Source (2002-12) into Target (2016-17) distributions in 3 regions using our method with ControlNet. The results illustrate our capacity to generate images across temporal discrepancies.

FMoW-WILDS [41] is a modified version of the Functional Map of the World [12] dataset. It includes 0.5 million RGB satellite images labeled with 62 land use categories, with domains defined by their captured years and geographic regions spanning Asia, Africa, Americas, Oceania and Europe. For this paper, we focused on the domain adaptation performance across different time periods within three regions: Asia, Americas, and Oceania. Specifically, within each region, the source and target domain refer to the satellite images taken between 2002-12 and 2016-17. We randomly sampled 64 images from each domain to calculate the gap. For each land use category, DoGE generated 64 images using ControlNet (*e.g.*, Fig. 6) with edit strength mean $c = 1.3$, accounting for approximately 10% increase in the dataset size. Tab. 3 shows that DoGE leads to higher performance than baselines in all experiments.

4.3.2 Segmentation Task

Besides classification problems, our method is also generally applicable to other computer vision tasks. To illustrate the versatility and generality of DoGE, we demonstrate our capability to improve cross-domain segmentation problems.

Method	Test Accuracy (%)
DAFormer [31]	48.2
DAFormer + DATUM [3]	56.4
DAFormer + DoGE (Ours)	**57.3**

Table 6. **GTA5 → Cityscapes cross-domain segmentation.** We used DAFormer as our UDA baseline. DATUM and DoGE are target data generators applied on top of DAFormer. Our performance is at par with DATUM while exempt from any training.

In this experiment, we chose GTA5 [60] as our source domain and Cityscapes [13] as our target domain. Using the full GTA5 dataset and 20 unlabeled images from the Cityscapes, we generated synthetic data in Fig. 5 with and without ControlNet. We evaluated our synthetic generation under the scope of UDA. As baselines, we chose DAFormer [31], a UDA segmentation method, and DATUM combined with DAFormer. Similar to DATUM, we evaluated our method on top of DAFormer, *i.e.*, expanding the unlabeled data available to DAFormer. Tab. 6 shows DoGE is able to achieve at-par performance with DATUM. Moreover, DATUM requires fine-tuning a Stable Diffusion model while ours is inference-only in a plug-and-play fashion.

4.4. Ablation Studies

4.4.1 Generation Quality

The usefulness of our synthetic data is directly dependent on the generation quality. We focus on two aspects to assess the generated images: the FID score [27] for image quality, and the t-SNE [73] for distribution alignment.

The exploration was conducted under our DomainNet Real→Painting experiments. Tab. 7 shows that under FID metrics with respect to DomainNet Painting images, our generation achieved the best quality with respect to the Painting data from DomainNet. To visualize the distribution alignment of our generation, we plotted the t-SNE graph of source (Real domain), target (Painting domain), and our

Data Source	FID Score (↓)
Source Data	30.98
DA-Fusion [71]	40.20
DATUM [3]	219.00
DoGE (Ours)	24.86
DoGE w/ ControlNet (Ours)	**18.25**

Table 7. **FID scores against the DomainNet painting images.** We evaluated the FID scores against the DomainNet Painting samples on the DomainNet Real images, the synthetic data from [71] and [3], and our generations. The table shows that our synthesis achieved the best FID score among the baselines.

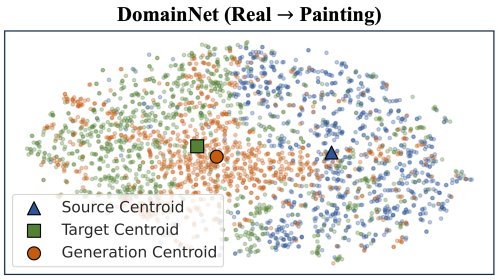

Figure 7. **The t-SNE plots of the source, target, and generated data.** In our DomainNet Real → Painting experiment, we drew a t-SNE plot to visualize distributions of source, target, and our generation. Our generation is well-aligned with the target distribution.

synthetic painting images in Fig. 7. It shows that our synthetic data are successfully augmented into the target distribution and away from the source distribution.

4.4.2 Domain Gap Embedding Editing Weights

One of the important hyper-parameters that impacts the generation is the edit strength scalar C defined in Sec. 3.2. In this section, we study the effect of different deterministic values for C visually to better understand the domain gap embeddings. To isolate the effect, we do not apply ControlNet in this experiment. As shown in Fig. 8, we conduct the exploration in two settings: face augmentation with eyeglasses as the distribution gap (top row), and object augmentation from the real domain to the sketch domain (bottom row). Starting with the source image (left-most column), we gradually increase the edit strength C and generate images at each different value for C. For the face, we set C to 0.5, 1.0, 1.5, 2.0 from left to right, and 1.0, 1.5, 2.0, 2.5 respectively for the airplane.

From Fig. 8, we observe that the magnitude of the edit strength impacts the extent of our augmentation. When $C \geq 2$ as shown in the right-most column, our pipeline adds two glasses on the face indicating an over augmentation. Meanwhile, for the airplane, the value of C influences whether the generation is a realistic sketch or a simple sketch. Hence the best choice of edit strength mean c should be assessed based on the kind of task in practice.

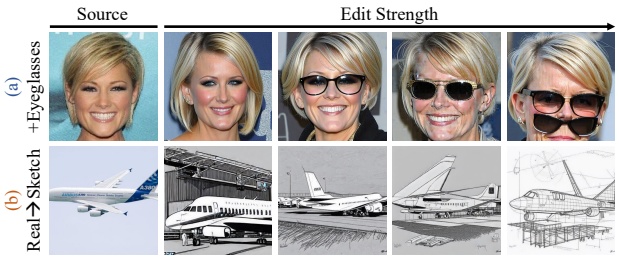

Figure 8. **Effect of increasing edit strength c.** We considered two source images under two tasks: (a) adding eyeglasses to faces and (b) converting real to sketch images. In each task, we generated images with gradually increasing edit strength. At the right end, we observe that two glasses are added $c = 2.0$ and the most sketchy airplane $c = 2.5$. As expected, the edit strength dictates the extent of emphasis on the distribution differences.

5. Conclusion and Future Works

This paper introduces DoGE, an innovative diffusion-based data augmentation technique designed to address cross-distribution challenges. Our method is distinguished by its accessibility, efficiency, and remarkable effectiveness. We utilize Domain Gap Embeddings, which capture distribution differences, as direct augmentations applied to source data embeddings. Our generative backbone, Stable UnCLIP, is leveraged to facilitate this process. It's worth noting that our pipeline operates without the necessity for training, relying exclusively on a minimal set of images from the target distribution to guide the augmentation process. The result is the generation of diverse and high-quality synthetic data, which significantly enhances test performance.

We showcase the versatility and effectiveness of our method across various problem settings. Notably, our approach not only excels at transferring styles but also introduces semantic augmentations according to distribution disparities. In comparison to other general data synthesis methods, we achieve the highest improvements across all tasks. We also highlight the adaptability of DoGE by evaluating its performance in a segmentation task, demonstrating competitive inference-only results compared to the state-of-the-art method, which requires training. Furthermore, we illustrate that our approach is compatible with and complementary to parallel strategies such as UDA and fine-tuning.

While our method boasts significant strengths, it does have certain limitations that merit further consideration. First and foremost, the expressiveness of the CLIP model's latent space can be constrained when confronted with domain gaps that CLIP is unfamiliar with. Additionally, while Stable UnCLIP proves effective in numerous real-world scenarios, it may face challenges in out-of-domain situations, such as medical X-ray imagery. Lastly, there is ample room for exploration in devising more effective training algorithms to maximize the utility of synthetic data.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. In *TMLR*, 2023. 2

[2] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023. 2

[3] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. One-shot unsupervised domain adaptation with personalized diffusion models. In *CVPRW*, 2023. 2, 3, 4, 6, 7, 8, 13

[4] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP*, 2020. 2

[5] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *WACV*, 2021. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[8] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. In *Neural Networks*, 2018. 2

[9] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *TMLR*, 2023. 2

[10] John Canny. A computational approach to edge detection. In *TPAMI*, 1986. 4

[11] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 2019. 6, 13

[12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 7

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 7

[14] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020. 1, 4, 5, 6

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[16] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023. 2

[17] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. 1901. 3

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[19] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 6, 13

[20] Rui Gong, Qin Wang, Dengxin Dai, and Luc Van Gool. One-shot domain adaptive and generalizable semantic segmentation with class-aware cross-domain transformers. *arXiv preprint arXiv:2212.07292*, 2022. 14

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[22] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023. 2

[23] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, 2023. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[25] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 2

[26] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 2

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 7

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPSW*, 2021. 2

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[30] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[31] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 7

[32] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 1

[33] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 5

[34] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. In *ICLR*, 2022. 2

[35] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, 2020. 6, 13

[36] Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker. Memsac: Memory augmented sample consistency for large scale domain adaptation. In *ECCV*, 2022. 6, 13

[37] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 2

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[39] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 2

[40] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 2

[41] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 1, 7

[42] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *CVPR*, 2022. 2

[43] Sebastian Nørgaard Llambias, Mads Nielsen, and Mostafa Mehdipour-Ghazi. Data augmentation-based unsupervised domain adaptation in medical imaging. In *arXiv preprint arXiv:2308.04395*, 2023. 1

[44] Justin Lo, Jillian Cardinell, Alejo Costanzo, and Dafna Sussman. Medical augmentation (med-aug) for optimal data augmentation in medical deep learning networks. In *Sensors*, 2021. 1

[45] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 6, 13

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[47] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De La Torre. Zero-shot model diagnosis. In *CVPR*, 2023. 2

[48] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *NeurIPS*, 2020. 14

[49] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2

[50] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[51] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[52] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. In *JAIR*, 2021. 2

[53] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS*, 2021. 4

[54] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[55] Duo Peng, Qiuhong Ke, Yinjie Lei, and Jun Liu. Unsupervised domain adaptation via domain-adaptive diffusion. In *ITIP*, 2023. 2

[56] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *CVPRW*, 2018. 2

[57] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. In *NeurIPS*, 2023. 2

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[60] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 7

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

[62] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[63] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the wilds benchmark for unsupervised adaptation. In *ICLR*, 2022. 1

[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[65] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 6, 13

[66] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1

[67] Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023. 2

[68] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *CVPRW*, 2023. 2

[69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[70] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: A synthetic driving dataset for continuous multi-task domain adaptation. In *CVPR*, 2022. 2

[71] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *ICLRW*, 2023. 2, 4, 5, 6, 8, 13

[72] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: Compositional structures in vision-language models. In *ICCV*, 2023. 3

[73] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. In *JMLR*, 2008. 7

[74] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023. 2

[75] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[76] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ICCV*, 2021. 2

[77] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023. 2

[78] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 4

[79] Jianhao Yuan, Francesco Pinto, Adam Davies, Aarushi Gupta, and Philip Torr. Not just pretty pictures: Text-to-image generators enable interpretable interventions for robust representations. In *arXiv preprint arXiv:2212.11237*, 2022. 2

[80] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *ICCV*, 2023. 2

[81] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

[82] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[83] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2

[84] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *ICMLW*, 2023. 2

[85] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Domainstudio: Fine-tuning diffusion models for domain-driven image generation using limited data. *arXiv preprint arXiv:2306.14153*, 2023. 3

[86] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[87] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *ICLR*, 2022. 2