

TTIDA: Controllable Generative Data Augmentation via Text-to-Text and Text-to-Image Models

**Yuwei Yin¹, Jean Kaddour², Xiang Zhang³, Yixin Nie⁴,
Zhenguang Liu⁵, Lingpeng Kong¹, Qi Liu¹**

¹ Department of Computer Science, University of Hong Kong; ² University College London

³ University of Alberta; ⁴ University of North Carolina at Chapel Hill; ⁵ Zhejiang University

{ywyin, lpk, liuqi}@cs.hku.hk; jean.kaddour.20@ucl.ac.uk;
xzhang23@ualberta.ca; yixin1@cs.unc.edu; zhenguangliu@zju.edu.cn

Abstract

Data augmentation has been established as an efficacious approach to supplement useful information for low-resource datasets. Traditional augmentation techniques such as noise injection and image transformations have been widely used. In addition, generative data augmentation (GDA) has been shown to produce more diverse and flexible data. While generative adversarial networks (GANs) have been frequently used for GDA, they lack diversity and controllability compared to text-to-image diffusion models. In this paper, we propose TTIDA (Text-to-Text-to-Image Data Augmentation) to leverage the capabilities of large-scale pre-trained Text-to-Text (T2T) and Text-to-Image (T2I) generative models for data augmentation. By conditioning the T2I model on detailed descriptions produced by T2T models, we are able to generate photo-realistic labeled images in a flexible and controllable manner. Experiments on in-domain classification, cross-domain classification, and image captioning tasks show consistent improvements over other data augmentation baselines. Analytical studies in varied settings, including few-shot, long-tail, and adversarial, further reinforce the effectiveness of TTIDA in enhancing performance and increasing robustness.¹

1 Introduction

Data augmentation is ubiquitous in the preprocessing procedure of various machine learning tasks, particularly those involving insufficient labeled data (Shorten and Khoshgoftaar, 2019). It provides multiple benefits, such as minimizing the costs associated with collecting and annotating data, alleviating concerns related to data scarcity and imbalance, and mitigating the deleterious effects of overfitting on model generalization. The efficacy of data augmentation is predicated on the extent

to which the augmented dataset approximates the underlying data distribution. Thus the optimal objective is to obtain a wide range of samples that reflect the natural distribution of richness and diversity in a given target object category.

Most conventional augmentation methods employed in computer vision rely on handcrafted transformations that utilize a restricted array of elementary invariances, such as rotation, cropping, and color adjustment (Shorten and Khoshgoftaar, 2019; Mikołajczyk and Grochowski, 2018; Fawzi et al., 2016). These transformations are pre-specified and applied uniformly across the entire dataset, which may not be optimal for different types of data or scenarios. Such a limitation motivates the need for more advanced and flexible approaches to augment data effectively in the context of various computer vision tasks. To obtain more diverse images, prior studies (Antoniou et al., 2017; Bowles et al., 2018) on generative data augmentation (GDA) aim to approximate the data distribution of the observed image dataset using generative adversarial networks (Goodfellow et al., 2020) (GANs). However, each category needs a separate GAN model to be trained, which is inflexible and incurs considerable training expenses. Besides, the training process of GANs is notoriously unstable, especially when the training set is small. It also brings the problem of mode collapse, which produces images with less diversity.

In this paper, we aim to address the limitations of existing GDA methods for vision tasks by exploring publicly accessible text-to-image (T2I) models, which generate images according to the input text. These models are based on diffusion models (Sohl-Dickstein et al., 2015; Rombach et al., 2022) and have been trained with large-scale text-image pairs obtained from the Web. Hence T2I models can generate a variety of photo-realistic images that are conditioned on diverse text descriptions. The benefits of utilizing these models are manifold. 1)

¹<https://github.com/YuweiYin/TTIDA>

they provide a language interface that enables flexible control and generation of desired images; 2) they are domain-agnostic and capable of generating extensive data with high diversity; 3) they serve as a versatile tool to synthesize high-quality data for vision tasks in various scenarios, including in-domain, cross-domain, long-tail, and so forth.

In the field of natural language processing, Edwards et al. (2021) have utilized the large pretrained text-to-text (T2T) model GPT-2 (Radford et al., 2019) as a method of data augmentation for text classification tasks. Building on this concept, we propose TTIDA (Text-to-Text-to-Image Data Augmentation) to leverage the generative capabilities of large-scale pretrained text-to-image (T2I) and text-to-text (T2T) models for the purpose of data augmentation. Specifically, we first fine-tune a T2T model, such as GPT-2 and T5 (Raffel et al., 2020), on a diverse set of image captions. Then, for each object category, the label text is inputted to the T2T model to obtain a more detailed description of the object of our desire. Following this, a T2I model, such as GLIDE (Nichol et al., 2022), is employed to generate multiple photo-realistic images of the object, which are conditioned on either the original label text or the more detailed description. The synthetic images thus produced are utilized to augment the original image dataset. Using detailed descriptions as prompts for the T2I model is beneficial as label text typically contains only one or two words and is, therefore, typically simplistic. Therefore, the use of T2T models brings controllability to the generated images, and the diversity of generated text also guarantees the diversity of generated images.

To evaluate the efficacy of our approach, a series of experiments are conducted on various tasks, including (1) image classification with distinct scenarios such as balanced, long-tail, and adversarial data settings, (2) cross-domain image classification, and (3) image captioning. The experimental results on CIFAR (Krizhevsky et al., 2009), Office (Saenko et al., 2010; Venkateswara et al., 2017), and MS COCO (Lin et al., 2014) benchmarks substantiate the consistent performance enhancement of our method compared to other data augmentation baselines. Remarkably, we observe that the superiority of our approach is more notable in instances of scarce data or diverse data domains. Furthermore, we show that TTIDA is capable of enhancing model robustness in the face of adversarial attacks.

Our contributions are summarized as follows.

- We propose TTIDA, a novel approach combining the generation power of large-scale pretrained text-to-text (T2T) and text-to-image (T2I) models for data augmentation in a controllable and flexible way.
- We demonstrate the efficacy of TTIDA in enhancing the model performance on in-domain classification, cross-domain classification, and image captioning benchmarks.
- The analytical studies in varied settings, including few-shot, long-tail, and adversarial, further reinforce the effectiveness and robustness of TTIDA.

2 Related Work

2.1 Traditional Image Transformations

The application of conventional augmentation techniques for computer vision has been widely acknowledged and validated to be effective (Perez and Wang, 2017). These techniques typically rely on manually crafted transformations that exploit a restricted set of elementary invariances (Goodfellow et al., 2016), including but not limited to cropping, rotation, and flipping (Shorten and Khoshgoftaar, 2019). The mere inclusion of a greater number of image augmentation techniques does not invariably culminate in an improvement in performance since certain methods may demonstrate a sensitivity to the selection of image augmentations (Grill et al., 2020).

2.2 Generative Data Augmentation

Antoniou et al. (2017) propose to use image-conditional Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) to generate within-class images conditioned on a source domain. Similarly, Bowles et al. (2018) use GANs to augment data for brain segmentation tasks. In addition to GANs, some researchers utilize the pretrained language model GPT-2 (Radford et al., 2019) as a method of data augmentation for multiple text classification tasks Edwards et al. (2021) and commonsense reasoning benchmarks (Yang et al., 2020). Besides generative approaches, another line of work tries to learn advanced augmentation strategies using a fixed subset of augmentation functions. Cubuk et al. (2019, 2020) use reinforcement learning to automatically find a dataset-specific augmentation policy which is proved to be effective on the downstream tasks.

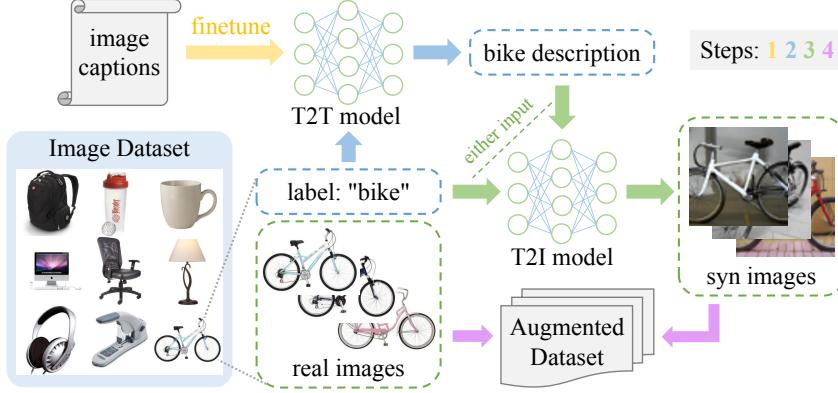


Figure 1: **Overview of TTIDA** (Text-to-Text-to-Image Data Augmentation). Arrows in different colors represent different steps. **Step 1**: finetune the text-to-text (T2T) model using text captions of images. **Step 2**: input the label text, i.e., “bike”, to the T2T model to produce a caption-like description of bikes. **Step 3**: input either the original label text or the generated description into the text-to-image (T2I) model to generate high-quality synthetic (syn) images. **Step 4**: combine the real images from the original dataset with the augmented images for model training.

3 Method

In this section, we elaborate on our TTIDA approach to generative data augmentation and the motivations behind our framework.

3.1 Overview

Figure 1 shows the overview of our data augmentation method, where arrows in different colors denote different pipeline steps. For each object category, i.e., bike in Figure 1, we input the label text “bike” to the T2I model such as GLIDE (Nichol et al., 2022) to generate multiple photo-realistic images of this object (**Step 3**). Then we combine the real images from the original dataset with the generated synthetic images together (**Step 4**). The augmented dataset is directly used for model training. Usually, the label text is a word or short phrase. To automatically obtain a finer prompt for the T2I model, we can first input the label text to a text-to-text (T2T) generative model finetuned with image captions (**Step 1**) to produce a longer object description (**Step 2**), e.g., “a white bike near the wall”. Step 1 and Step 2 are optional since the T2I model can still generate high-quality images with the label text input. Yet the T2T model can produce precise or personalized object descriptions with a richer context, increasing the diversity of synthetic images to a large extent.

3.2 Formulation

We denote the training set of an image classification dataset as $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$ that has n different categories. The i -th category is $\mathcal{D}_i = \{l_i, \mathcal{X}_i\}$, where l_i is the label text, and $\mathcal{X}_i = \{x_1, \dots, x_m\}$

are m real images in this category. **Firstly**, we finetune the T2T model $t2t(\cdot)$ on the caption corpus with the language modeling loss (next-token prediction) conditioned on the corresponding prompt p , i.e., to maximize:

$$\sum_u \sum_v \log p(t_v^u | t_{<v}^u; p^u), \quad (1)$$

where u is the number of captions in the training set of image captioning datasets, and v is the token index of the u -th caption. $t_{<v}^u$ denotes all the previous generated tokens before the v -th token. The prompt p^u of the u -th caption includes the caption’s entities.

Then, we input the category label l_i to the T2T model to obtain the description sentence:

$$d_i = t2t(l_i). \quad (2)$$

After that, the T2I model $t2i(\cdot)$ uses either l_i or d_i as the input prompt to generate plenty of diverse synthetic images $\{\hat{\mathcal{X}}_i = \hat{x}_i^j\}_{j=1}^G$ of the same object as depicted in l_i or d_i , where G is the number of synthetic images:

$$\text{either } \hat{x}_i^j = t2i(l_i) \quad (3)$$

$$\text{or } \hat{x}_i^j = t2i(d_i). \quad (4)$$

Lastly, we merge the original data of the i -th category $\mathcal{D}_i = \{l_i, \mathcal{X}_i\}$ with the generated images $\hat{\mathcal{X}}_i$ to construct the augmented category \mathcal{D}_i^{aug} :

$$\mathcal{D}_i^{aug} := \{l_i, \mathcal{X}_i, \hat{\mathcal{X}}_i\} = \mathcal{D}_i \cup \hat{\mathcal{X}}_i. \quad (5)$$

Repeat this process for every category \mathcal{D}_i in the dataset \mathcal{D} , we can have a augmented dataset $\mathcal{D}^{aug} = \{\mathcal{D}_1^{aug}, \dots, \mathcal{D}_n^{aug}\}$.

3.3 Text-to-Text Models

We use Generative Pre-Training (GPT-2) (Radford et al., 2019) as the T2T model. Specifically, we adopt the basic GPT2LMHeadModel² and set the maximal sentence length as 20 and the beam size as 5 for beam search. The GPT model is finetuned with the language modeling loss, and Adam optimizer (Kingma and Ba, 2015) for 5 epochs using all the captions in the training set of MS COCO 2015 Image Captioning Task (Lin et al., 2014).

When fine-tuning GPT-2, the prompt p^u in Equation 1 is a template containing all entity tokens $\{e_i^u\}_{i=1}^n$ extracted from the u -th caption sentence beforehand. We prepend the template “*Write an image description with keywords including $e_1^u, e_2^u, \dots, e_n^u$:*” to the original caption sentence. Then we feed the prompted caption into the GPT-2 model to finetune GPT-2 using the language modeling loss. In this way, by editing the entity words of input prompts when generating image captions using the fine-tuned GPT-2 model, we can flexibly control the generated sentences and thus can modify the contents of the generated images by the T2I model.

3.4 Text-to-Image Models

We adopt Guided Language to Image Diffusion for Generation and Editing (GLIDE) (Nichol et al., 2022) as our T2I model. Unlike other text-to-image models that mainly focus on the generation of pictures with different artistic styles, such as DALL-E (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022), GLIDE aims to generate photo-realistic pictures from the input prompt text. We feed the T2I model with image description sentences produced by the T2T model to generate synthetic images as data augmentation in a more controllable manner.

4 Experimental Setup

In this section, we elaborate on all the experimental setups, including task introduction (4.1), datasets (4.2), data augmentation (4.3), our method and baseline methods (4.4), backbone models (4.5). See Appendix C for detailed training details.

4.1 Tasks

To illustrate the versatility of our method, we tackle a diverse set of computer vision tasks. We now explain on a case-by-case basis how these settings can demonstrate the benefits of TTIDA.

In-domain Image classification In this task, the domain of the training set is the same as that of the test set. We consider a balanced data settings: Each category has the same number of images. Other settings such as Few-shot, long-tail, and adversarial will be discussed in Section 6.

Cross-domain Image classification The model trains on the images of one domain and tests on those of another. For example, a cross-domain dataset \mathcal{D} has K domains $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$, and each domain \mathcal{D}_k contains the same n categories $\{\mathcal{C}_1^k, \dots, \mathcal{C}_n^k\}$. The images in the i -th category \mathcal{C}_{src}^i of the source domain \mathcal{D}_{src} and those in the i -th category \mathcal{C}_{tgt}^i of the target domain \mathcal{D}_{tgt} denote the same object. Images of different domains are visibly different in some aspects, such as angle and position, tone of color, and background features. Hence, experimental results on this task will demonstrate the benefits of generating diverse images and thus verify the domain-agnostic trait of our method.

Image Captioning In this task, models need to generate a caption sentence that describes the input image accurately. By testing the performance of TTIDA on this task, we show the advantages of combining the power of autoregressive language models (T2T models) and T2I models to generate reasonable (text, image) pairs for improving the image-to-text generation (captioning) ability.

4.2 Datasets

In-domain Image classification We conduct in-domain image classification experiments on two datasets. CIFAR-100 contains 100 different distinguishable classes and a smaller size of training samples—500 images—in each class. We use the classification accuracy as the evaluation metric.

Cross-domain Image classification We adopt two cross-domain datasets to measure the effectiveness of TTIDA. 1) The Office-31 dataset (Saenko et al., 2010) contains 31 object categories in three domains: *Amazon*, *DSLR* and *Webcam*. 2) Office-Home (Venkateswara et al., 2017) is another benchmark dataset for domain adaptation containing 4 domains including *Art*, *Clipart*, *Product*, and *Real-World*, where each domain has 65 categories.

Image Captioning We use the image captioning dataset of the Microsoft COCO (common objects in context) 2015³ Image Captioning Task. MS COCO

²<https://huggingface.co/gpt2>

³<https://cocodataset.org/#captions-2015>

Captions (Lin et al., 2014) contains over one and a half million captions describing over 330,000 images. For the training and validation images, five independent human-generated captions are provided for each image.

4.3 Data Augmentation

Synthetic Images for CIFAR The images in the CIFAR dataset have the size of 32×32 ; simply specifying the generation with such low resolution from GLIDE will largely decrease the quality of images. Instead, we generate images of size 256×256 with more details and then perform a resizing. For CIFAR-100, 500 images are produced by the GLIDE model. In total, we provide 50000 additional training samples for these two datasets.

Synthetic Images for Office Unlike CIFAR, the original images in the Office-31 and Office-Home datasets vary from low resolution to high resolution. Besides, these images are not in the same shape, so CDTrans performs a series of image transformations. Our synthetic images are of size 256×256 initially and then perform the same transformations as in CDTrans. For each category in each domain, the number of synthetic images generated by T2I models is the same as that of Office datasets.

Synthetic Images for MS COCO We generate synthetic images using the training set of COCO captions as augmentation. In addition, we extract entities from COCO captions using CoreNLP⁴ and Natural Language Toolkit (NLTK)⁵, and then feed the T2I model with the entities or synthetic captions generated by T2T models using these entities.

4.4 Our Method and Baseline Methods

TTIDA We use the T2T and T2I models as described in Section 3.3 and 3.4. Specifically, for all experiments, we use the basic model settings of the public GLIDE model⁶ and follow the standard generation procedure: we use the base GLIDE model to generate 64×64 images, and then feed them into the upsample model to obtain high-quality images of 256×256 resolution. Then we perform a resizing to match the image size of different datasets. Except for resizing and image normalization, no other image transformations are adopted.

⁴<https://stanfordnlp.github.io/CoreNLP/>

⁵<https://www.nltk.org/>

⁶<https://github.com/openai/glide-text2im>

Traditional Image Transformations The incorporation of additional image augmentation techniques does not inevitably lead to improved performance, as certain methods may be sensitive to the choice of augmentations (Grill et al., 2020). Thus we adopt the image transformation procedure⁷ employed in SimCLR (Chen et al., 2020a) and MoCo (He et al., 2020; Chen et al., 2020b, 2021), which has been validated for its efficacy.

GAN-based Generative Data Augmentation

To contrast our TTIDA approach with previous generative data augmentation methods, we employ three representative GAN models, namely DC-GAN (Radford et al., 2016), CycleGAN (Zhu et al., 2017), and StyleGAN (Karras et al., 2021b, 2020, 2021a), to augment images. More specifically, we adhere to the original implementations of DC-GAN⁸, CycleGAN⁹, and StyleGAN¹⁰ for GAN training and generation.

4.5 Backbone Models

In-domain Image Classification For all experiments on CIFAR, we use the standard ResNet-101 architecture (He et al., 2016) as the backbone¹¹.

Cross-domain Image Classification We adopt the state-of-the-art model CDTrans (Xu et al., 2022) as the baseline model. We follow the implementation of CDTrans¹² and only add a data processing module for incorporating our synthetic images into the original source-domain dataset.

Image Captioning We use the state-of-the-art model mPLUG (Li et al., 2022) as our baseline and follow the implementation of mPLUG¹³. Similarly, we add a data processing block for integrating our augmented data into the COCO captioning dataset.

5 Results

In this section, we report all experimental results w.r.t. the settings described in Section 4.

⁷<https://github.com/facebookresearch/moco-v3>

⁸<https://github.com/pytorch/examples/tree/main/dcgan>

⁹<https://github.com/junyanz/CycleGAN>

¹⁰<https://github.com/NVlabs/stylegan3>

¹¹<https://pytorch.org/vision/main/models/generated/torchvision.models.resnet101.html>

¹²<https://github.com/CDTrans/CDTrans>

¹³<https://github.com/alibaba/AliceMind/tree/main/mPLUG>

CIFAR-100	+ 20%	+ 50%	+ 100%	+ Max
Img Trans DA	+0.3%	+0.4%	+0.4%	+0.5%
DCGAN DA	+0.4%	+0.5%	+0.7%	+1.0%
CycleGAN DA	+0.5%	+0.7%	+1.0%	+1.2%
StyleGAN DA	+0.7%	+0.9%	+1.2%	+1.4%
TTIDA (label)	+1.1%	+1.8%	+2.3%	+2.7%
TTIDA (desc.)	+1.3%	+2.1%	+2.6%	+3.0%

Table 1: **Classification accuracy on CIFAR-100.** We report accuracy improvements when adding synthetic images generated by different models described in Section 4.4. “+Max” denotes the highest score using 200%, 300%, 400%, or 500% synthetic images.

Office-31	A→D	A→W	D→A
Before Tuning	97.0%	96.7%	81.1%
w/o Syn Data	97.4%	96.8%	81.3%
w/ Syn Data	98.0%	97.1%	81.6%
Office-31	D→W	W→A	W→D
Before Tuning	99.0%	81.9%	100%
w/o Syn Data	99.0%	82.0%	100%
w/ Syn Data	99.1%	82.2%	100%

Table 2: **Target-domain classification accuracy in every direction on the Office-31 dataset.** Office-31 has three domains, namely Amazon (A), DSLR (D), and Webcam (W). “Before Tuning” stands for the test scores of the best checkpoints of the state-of-the-art model CDTrans. “w/ Syn Data” and “w/o Syn Data” represent the results after finetuning with and without synthetic images generated by TTIDA respectively.

5.1 In-domain Image Classification

The experimental results of image classification on the CIFAR-100 dataset are shown in Table 1. TTIDA outperforms all baselines on each synthetic ratio, demonstrating the effectiveness of our method in boosting the image classification performance.

5.2 Cross-domain Image Classification

We continue training the best CDTrans checkpoints for 50 epochs on every Office domain adaption direction with and without our synthetic images. The target-domain classification results on Office-31 and Office-Home datasets are reported in Table 2 and Table 3, respectively. We observe that fine-tuning consistently enhances classification accuracy for all directions, especially when training with our augmented data. The results verify the effectiveness of TTIDA in improving model performance on cross-domain image classification tasks.

5.3 Image Captioning

We train the base mPLUG (Li et al., 2022) model (mplug.en.base) based on pre-trained CLIP (Rad-

Art	A→C	A→P	A→R	Avg A
Before Tuning	68.8%	85.0%	86.9%	80.23%
w/o Syn Data	68.9%	85.4%	87.1%	80.47%
w/ Syn Data	69.2%	85.7%	87.6%	80.83%
Clipart	C→A	C→P	C→R	Avg C
Before Tuning	81.5%	87.1%	87.3%	85.30%
w/o Syn Data	81.8%	87.2%	87.4%	85.47%
w/ Syn Data	82.2%	87.5%	87.4%	85.70%
Product	P→A	P→C	P→R	Avg P
Before Tuning	79.6%	63.3%	88.2%	77.03%
w/o Syn Data	79.7%	64.5%	88.3%	77.50%
w/ Syn Data	80.1%	65.9%	88.5%	78.17%
RealWorld	R→A	R→C	R→P	Avg R
Before Tuning	82.0%	66.0%	90.6%	79.53%
w/o Syn Data	82.6%	66.1%	90.7%	79.80%
w/ Syn Data	82.8%	66.4%	90.9%	80.03%

Table 3: **Target-domain classification accuracy in every direction on the Office-Home dataset.** Office-Home has four domains, namely Art (A), Clipart (C), Product (P), and RealWorld (R). “Before Tuning” stands for the test scores of the best CDTrans checkpoints. “w/ Syn Data” and “w/o Syn Data” represent the results after finetuning for 50 epochs with and without synthetic images generated by TTIDA, respectively.

#Data + Syn Ratio	BLEU4	ROUGE-L	CIDEr-D
5000 + 100%	+3.2%	+1.0%	+2.4%
10000 + 100%	+1.9%	+2.1%	+3.0%
50000 + 100%	+1.5%	+1.3%	+5.2%
100000 + 100%	+0.5%	+0.2%	+1.0%
200000 + 100%	+0.7%	+0.3%	+2.1%

Table 4: **Test scores of mPLUG model finetuned on MS COCO 2015 Image Captioning dataset.** We report BLEU, ROUGE, and CIDEr scores with (+ 100%) and without (+ 0) synthetic images generated by TTIDA.

ford et al., 2021) model (ViT-B-16) for 5 epochs on the COCO image captioning dataset of different training size. The model is evaluated using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) metrics. Specifically, we adopt the evaluator implementation (Chen et al., 2015) to calculate BLEU4, sentence-level ROUGE-L, and CIDEr-D scores¹⁴.

Table 4 compares the test scores of the finetuned mPLUG model with and without the synthetic data generated by TTIDA. As the results show, TTIDA can further boost the model performance on different evaluations under different data size settings.

¹⁴<https://github.com/tylin/coco-caption>

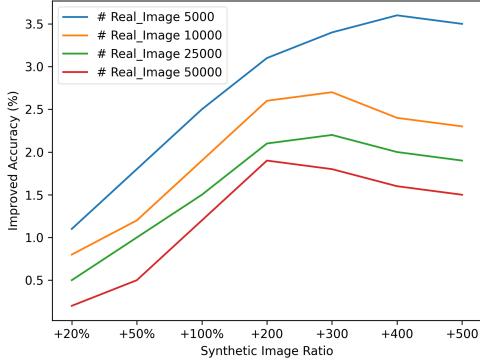


Figure 2: Results on the CIFAR-100 benchmark under the few-shot setting.

6 Analysis

In this section, we conduct analytic studies to better understand how our proposed framework contributes to the model performance.

6.1 Synthetic Images of Different Ratios

Commonly, data augmentation is more useful in low-resource settings than in high-resource ones. To verify the idea, we proportionately adjust the scale of the training set to create high- and low-resource contexts. Augmentation of the original training set with synthetic images of varying ratios is executed with the aim of determining the optimal conditions for the application of TTIDA. The experimental results of image classification on the CIFAR-100 dataset are shown in Figure 2. We report the classification accuracy (%) on the CIFAR datasets with synthetic training images of different proportions to the total number of original images.

It reveals that the incorporation of a greater number of synthetic images leads to a discernible reduction in classification error across all cases. Additionally, it is observed that the efficacy of the proposed method is particularly prominent in situations where the quantity of original CIFAR training images is meager. These findings suggest that the augmentation strategy proposed in this work holds particular promise in cases where the original dataset is characterized by limited resources. To further bolster this claim, the efficacy of TTIDA is evaluated on synthetic long-tail CIFAR datasets.

6.2 Training on Long-tail Datasets

We test our method on the synthetic long-tail CIFAR subset, where each category has a different

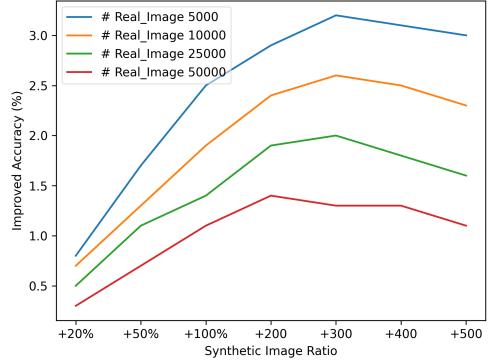


Figure 3: Results on the Long-tail CIFAR-100 benchmark under the few-shot setting.

Dataset	+ 100% Syn	+ Adv	Acc
CIFAR-100	✓	✓	-1.8%
CIFAR-100	✓	✓	-0.7%

Table 5: Classification accuracy on CIFAR-100 with adversarial training images. We set the number of original training images to 10k and report the results with and w/o synthetic (Syn) and adversarial (Adv) data added to the training set.

number of images. We construct the Long-tail CIFAR subset $\mathcal{D}^{lt} = \{\mathcal{D}_1^{lt}, \dots, \mathcal{D}_n^{lt}\}$ from the original balanced CIFAR dataset $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$ that has as n different categories, where the i -th sub-category \mathcal{D}_i^{lt} has i/n of the images in \mathcal{D}_i .

We report the improvement of classification accuracy (%) after using synthetic images generated by TTIDA on the long-tail CIFAR-100 dataset. As shown in Figure 3, the experimental results are consistent with the findings in Figure 2. Thus we conclude that the effectiveness of TTIDA in improving the accuracy of image classification, especially in low-resource cases, is further verified.

6.3 Training with Adversarial Images

We conduct experiments after adding adversarial images to the training set. For each class, we manually collect such images from the Internet, e.g., images with unusual styles (Appendix A). This experiment will test whether a diverse set of synthetic data helps to make the classifier more robust.

As shown in Table 5, we compare the differences in classification accuracy before and after adding the adversarial images to the test set. The model trained with augmented data performs 2.80% better, demonstrating that it boosts model robustness to unusual images.

Dataset	label	description	Acc
CIFAR-100	✓		+1.7%
CIFAR-100		✓	+1.9%
Dataset	label	description	Avg Acc
Office-31	✓		+0.25%
Office-31		✓	+0.35%
Dataset	label	description	BLEU4
MS COCO	✓		+1.5%
MS COCO		✓	+2.1%

Table 6: **Comparison of T2I prompts.** We compare the test scores (CIFAR-100, Office-31, MS COCO) after augmenting with generated images with the original label text or descriptions generated by TTIDA.

6.4 Prompts Generated by T2T Models

As shown in Figure 1, we can prompt the T2I model by either just a short label name or longer descriptions about the to-be-generated image. In this ablation, we compare the model performance between using the original label text or caption-like descriptions as prompts for the T2I model.

Table 6 shows that using rich descriptions to generate images results in a better improvement than using label text, especially on MS COCO benchmark. We hypothesize that using descriptions is more beneficial to cross-domain tasks and the large-scale COCO task since sentences produced by T2T models have richer context so that the T2I model can generate more diverse images. By contrast, augmented data with too much diversity is less effective for the in-domain classification task CIFAR.

6.5 Finetuning T2I Models

To further explain the observations and verify the hypothesis in the former section, we finetune the T2I model on the training set of CIFAR-100 so that the synthetic images look more similar to the original images. Specifically, we follow the open-sourced practice¹⁵ to finetune GLIDE using the training set of MS COCO image captioning dataset. Table 7 demonstrates that finetuning the T2I model brings further gains as expected.

7 Future Work

7.1 Multimodal Back-Translation

In machine translation, back-translation (Sennrich et al., 2016) is widely used as a powerful data augmentation method. It performs a

Dataset	label	description	finetune	Acc
CIFAR-100	✓			+1.5%
CIFAR-100	✓		✓	+2.0%
CIFAR-100		✓		+1.9%
CIFAR-100		✓	✓	+2.3%

Table 7: **Comparison of finetuning T2I models.** We compare the classification accuracy between using the original label text or descriptions on CIFAR-100 image classification task. The text-to-image model is finetuned on the CIFAR-100 training set if “finetune” is checked.

source→target→source translation process to augment the text data of the source language. Similarly, a multimodal back-translation can be conducted using a text-to-image model like GLIDE and an image-to-text model like the model used in image captioning tasks. Given available high-quality generators, we can generate extra text and image data by performing text→image→text and image→text→image back-translation respectively.

7.2 Spurious Correlations

Spurious correlations are correlations between image features and certain target classes where the features do not cause the latter (Kaddour et al., 2022). They naturally occur in many datasets (Neuhaus et al., 2022; Vasudevan et al., 2022; Lynch et al., 2023), but a model relying on them becomes problematic when faced with images where these correlations do not hold. Given the flexibility of TTIDA, we believe a promising direction is to augment training sets with images containing non-spurious features. This could be an attractive alternative for mitigating a model’s reliance on spurious features compared to the current paradigm of downsampling majority groups (Idrissi et al., 2022; Schwartz and Stanovsky, 2022; Arjovsky et al., 2022).

8 Conclusion

In this work, we propose TTIDA, a data augmentation scheme using large-scale pre-trained text-to-image and text-to-text models. We focus on generating photo-realistic images based on (i) concise label text and (ii) longer description prompts. Experimental results conducted for image classification and image captioning tasks under different settings demonstrate the effectiveness and robustness of the proposed augmentation method. Furthermore, we ablate some of our method’s components and conclude with future work directions.

¹⁵<https://github.com/afiaka87/glide-finetune>

9 Limitations

Despite the effectiveness of our method on the tasks above, this data augmentation approach can be applied to more machine learning applications, especially in low-resource situations. More experiments on different tasks with a wide range of hyper-parameter searching can further solidify our conclusions. Besides, it is imperative to explore filtering strategies to avoid inappropriate data being added to the training set, e.g., outlier samples that are too dissimilar to the training set distribution according to some distance measure.

10 Ethics Statement

As described in (Nichol et al., 2022, Section 6), the original GLIDE model generates fake but realistic images with possible disinformation or biases introduced by the data it was trained on. However, we believe that such ethical concerns can be addressed by appropriate data filtering, as suggested by Nichol et al. (2022).

References

- Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. 2017. [Data augmentation generative adversarial networks](#). *CoRR*, abs/1711.04340.
- Martín Arjovsky, Kamalika Chaudhuri, and David Lopez-Paz. 2022. [Throwing away data improves worst-class error in imbalanced classification](#). *CoRR*, abs/2205.11672.
- Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger N. Gunn, Alexander Hammers, David Alexander Dickie, María del C. Valdés Hernández, Joanna M. Wardlaw, and Daniel Rueckert. 2018. [GAN augmentation: Augmenting training data using generative adversarial networks](#). *CoRR*, abs/1810.10863.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. [A simple framework for contrastive learning of visual representations](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, Virtual Event. PMLR.
- Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020b. [Improved baselines with momentum contrastive learning](#). *CoRR*, abs/2003.04297.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. [Microsoft COCO captions: Data collection and evaluation server](#). *CoRR*, abs/1504.00325.
- Xinlei Chen, Saining Xie, and Kaiming He. 2021. [An empirical study of training self-supervised vision transformers](#). In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9620–9629, Montreal, QC, Canada. IEEE.
- Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2019. [Autoaugment: Learning augmentation strategies from data](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 113–123, Long Beach, CA, USA. Computer Vision Foundation / IEEE.
- Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2020. [RandAugment: Practical automated data augmentation with a reduced search space](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 3008–3017, Seattle, WA, USA. Computer Vision Foundation / IEEE.
- Aleksandra Edwards, Asahi Ushio, José Camacho-Collados, Hélène de Ribaupierre, and Alun D. Preece. 2021. [Guiding generative language models for data augmentation in few-shot text classification](#). *CoRR*, abs/2111.09064.
- Alhussein Fawzi, Horst Samulowitz, Deepak S. Turaga, and Pascal Frossard. 2016. [Adaptive data augmentation for image classification](#). In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 3688–3692, Phoenix, AZ, USA. IEEE.
- Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. 2016. [Deep Learning](#). Adaptive computation and machine learning. MIT Press, Cambridge, MA, USA.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2020. [Generative adversarial networks](#). *Commun. ACM*, 63(11):139–144.
- Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. [Bootstrap your own latent - A new approach to self-supervised learning](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33, pages 21271–21284, Virtual Event. NeurIPS.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA*,

June 13-19, 2020, pages 9726–9735, Seattle, WA, USA. Computer Vision Foundation / IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, Las Vegas, NV, USA. IEEE Computer Society.

Badr Youbi Idrissi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. 2022. Simple data balancing achieves competitive worst-group accuracy. In *1st Conference on Causal Learning and Reasoning, CLeaR 2022, Sequoia Conference Center, Eureka, CA, USA, 11-13 April, 2022*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351, Eureka, CA, USA. PMLR.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. Causal machine learning: A survey and open problems. *CoRR*, abs/2206.15475.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021a. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 852–863, Virtual Event. NeurIPS.

Tero Karras, Samuli Laine, and Timo Aila. 2021b. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116, Seattle, WA, USA. Computer Vision Foundation / IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, San Diego, CA, USA. ICLR.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. *University of Toronto*, 1.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7241–7259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, Zurich, Switzerland. Springer.

Aengus Lynch, Gbètondji J.-S. Dovonon, Jean Kadour, and Ricardo Silva. 2023. Spawrious: A benchmark for fine control of spurious correlation biases. *CoRR*, abs/2303.05470.

Agnieszka Mikołajczyk and Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, volume 1, pages 117–122, Swinoujscie, Poland. IEEE, Institute of Electrical and Electronics Engineers (IEEE).

Yannic Neuhaus, Maximilian Augustin, Valentyn Borейko, and Matthias Hein. 2022. Spurious features everywhere - large-scale detection of harmful spurious features in imagenet. *CoRR*, abs/2212.04871.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804, Baltimore, Maryland, USA. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318, Philadelphia, PA, USA. ACL.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

- Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, Virtual Event. PMLR.
- Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, San Juan, Puerto Rico. ICLR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685, New Orleans, LA, USA. IEEE.
- Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 213–226, Heraklion, Crete, Greece. Springer.
- Roy Schwartz and Gabriel Stanovsky. 2022. On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2182–2194, Seattle, WA, United States. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, volume 1, Berlin, Germany. The Association for Computer Linguistics.
- Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.
- Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265, Lille, France. JMLR.org.
- Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. 2022. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *CoRR*, abs/2205.04596.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575, Boston, MA, USA. IEEE Computer Society.
- Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5385–5394, Honolulu, HI, USA. IEEE Computer Society.
- Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. 2022. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, Virtual Event. OpenReview.net.
- Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Virtual Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1008–1025, Virtual Event. Association for Computational Linguistics.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251, Venice, Italy. IEEE Computer Society.

A Case Study

Original Images Figure 4 shows the original bike images of the *Art* and *Real-World* domain of Office-Home dataset. The image backgrounds and styles are visibly different in different domains.



(a) Bikes in Office-Home *Art*.



(b) Bikes in Office-Home *Real-World*.

Figure 4: Original bike images in the *Art* and *Real-World* domain of Office-Home dataset.

Our Synthetic Images Figure 5 shows the synthetic bike images generated by TTIDA. We can observe that the vanilla GLIDE model is prone to generate images resembling photo-realistic images in the real world.

Adversarial Images Figure 6 shows the adversarial images we collected from the Internet. These images have unusual backgrounds or strange styles, so they break the spurious correlations between the content and background in the original dataset.

B Data Statistics

Table 8 lists the statistics of the datasets in in-domain (CIFAR) and cross-domain (Office) image classification tasks.



Figure 5: Synthetic bike images generated by TTIDA.



Figure 6: Adversarial images collected from the Internet.

C Training Details

In-domain Image Classification The standard ResNet-101 model is trained from inception for a duration of 200 epochs on a single NVIDIA A40 GPU. To evaluate the model’s efficacy, a holdout validation set is used, which is randomly sampled from every category of the training set by 20 percent and is assessed after each epoch. Post-training, the optimal checkpoint is determined by selecting the best-performing model on the validation set. The selected checkpoint is then utilized to evaluate the accuracy of the model on the test set. For each setting, we repeat the process three times with three different random seeds $\{7, 17, 42\}$ and report the average test score.

For all classification experiments on CIFAR, the loss function is cross-entropy, and the batch size is 128. We use the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005. We perform a multi-step learning rate scheduler with gamma as 0.2 and milestones as $\{60, 120, 160\}$. In

Dataset (<i>Domain</i>)	# total	# class	# per class
CIFAR-100	50000	100	500
Office-31 (<i>Amazon</i>)	2817	31	91
Office-31 (<i>DSLR</i>)	498	31	16
Office-31 (<i>Webcam</i>)	795	31	26
Office-Home (<i>Art</i>)	2427	65	37
Office-Home (<i>Clipart</i>)	4365	65	67
Office-Home (<i>Product</i>)	4439	65	68
Office-Home (<i>Real-World</i>)	4357	65	67

Table 8: Statistics of image classification datasets. CIFAR-100 has 50000 images for training and extra 10000 images for testing. Office-31 dataset has 3 different domains and Office-Home dataset has 4 domains.

addition, the first 10 epochs are warm-up epochs with linearly increasing learning rates.

Cross-domain Image Classification We follow all the training settings of CDTrans ([Xu et al., 2022](#)) and finetune the published best checkpoint on the Office benchmark for 50 epochs.

Image Captioning Similarly, We follow all the training details of mPLUG ([Li et al., 2022](#)) and finetune the published best checkpoint on the MS COCO training set for 5 epochs.