Generative Data Augmentation for Few-Shot Domain Adaptation

Carlos E. López Fortín[a] and Ikuko Nishikawa[b]

Graduate Department of Information Science & Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan

Keywords: Domain-Adaptation, Diffusion-Model, Few-Shot, Data-Augmentation.

Abstract: Domain adaptation in computer vision focuses on addressing the domain gap between source and target distributions, generally via adversarial methods or feature distribution alignment. However, most of them suppose the availability of sufficient target data to properly teach the model domain-invariant representations. Few-shot scenarios where target data is scarce pose a significant challenge for their implementation in real-world scenarios. Leveraging fine-tuned diffusion models for synthetic data augmentation, we present Generative Data Augmentation for Few-shot Domain Adaptation, a model-agnostic approach to address the Few-shot problem in domain adaptation for multi-class classification. Experimental results show that using augmented data from fine-tuned diffusion models with open-source data sets can improve average accuracy by up to 3%, as well as increase per-class accuracy between 3% to 30%, for state-of-the-art domain adaptation methods with respect to their non-augmented counterparts, without requiring any major modifications to their architecture. This provides an easy-to-implement solution for the adoption of domain adaptation methods in practical scenarios.

1 INTRODUCTION

The ability for a model to learn representations from labeled data of a known distribution (source domain) and transfer that knowledge to another distribution (target domain) without additional supervision is known as Domain Adaptation (DA) (Goodfellow et al., 2014; Long et al., 2015; Ganin and Lempitsky, 2015). This has been a core challenge for the generalization of image classification and image segmentation models within the last few years (Liu et al., 2022). Most state-of-the-art models rely on techniques to partially or totally align the source and target distributions (Saito et al., 2020; Ganin et al., 2016; Yu et al., 2023), as well as methods to adversarially learn domain-invariant representations across domains (You et al., 2019; Yu et al., 2023). Despite the impressive performance achieved by these methods both in closed-set and open-set scenarios, however, most of them assume that enough target data is available during training, which is not always true in practice due to factors such as time availability, budgetary constraints, or technical limitations (Liu et al., 2022). This poses a serious obstacle for their adoption in real-world applications.

Few-shot scenarios in DA occur when the amount of target data available for model training is significantly less than the source data (Zhao et al., 2021; Liu et al., 2022). This can take place both at the individual class level (i.e. only some classes have limited samples) and at a general level (e.g. all classes from the target have limited samples). Some models have attempted to circumvent this issue by introducing prototype learning and additional adversarial learning components to leverage limited information from available samples (Zhao et al., 2021; Motiian et al., 2017). Others have attempted data augmentation through style-transfer from the target domain onto source-images or through data generation by diffusion models (Yang et al., 2021; Benigmim et al., 2023).

Inspired by the results of the DATUM model which utilized personalized diffusion models to address the One-shot scenario for image segmentation (Benigmim et al., 2023), we propose in this paper to leverage fine-tuned diffusion models to generate additional synthetic and diversified high-quality target data for Few-shot scenarios, which can be utilized with any DA model to improve their performance in multi-class classification. Our proposed method, Generative Data Augmentation for Few-Shot Domain Adaptation, significantly contributes to improving the model average and/or per-class accuracy of existing DA methods, and thus offers an interesting alterna-

[a] https://orcid.org/0000-0002-8727-7824
[b] https://orcid.org/0000-0003-4780-0155

López Fortín, C. and Nishikawa, I.
Generative Data Augmentation for Few-Shot Domain Adaptation.
DOI: 10.5220/0012338000003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 256-265
ISBN: 978-989-758-684-2; ISSN: 2184-4313

tive for the research of Few-shot domain adaptation. In addition, it offers a simple yet innovative solution for the adoption of state-of-the-art algorithms in real-world scenarios thanks to its model-agnostic approach, which, to our knowledge, has not been proposed before for multi-class classification.

In Section 2, we introduce related work for Domain Adaptation, Few-shot learning, and diffusion models for data generation. We continue Section 3 by defining the problem and describing the method for combining the results of fine-tuned diffusion models with DA algorithms. We discuss our results in Section 4 and provide a more detailed analysis in Section 5. Finally, we summarize our results in Section 6 and propose further research directions.

2 RELATED WORKS

2.1 Domain Adaptation

When a model is trained on a certain distribution of data, it is expected to perform similarly with new data with similar distributions. However, when the distribution is different, a domain gap is introduced, which drastically reduces its accuracy, and thus techniques to mitigate this difference are adopted (Liu et al., 2022; Long et al., 2015). Domain adaptation (DA) is a special class of transfer learning that focuses on minimizing the distribution discrepancy between different domains (You et al., 2019; Cao et al., 2019). This is generally achieved through adversarial techniques that allow the model to learn domain-invariant representations of classes from the source and target domains, as well as by partial or total alignment of data distributions at different stages of the training (You et al., 2019; Saito et al., 2020; Saito et al., 2020; Cao et al., 2018; Yu et al., 2023). Depending on where this alignment is carried out during training (input, latent, or output space), different results are obtained (Ganin and Lempitsky, 2015; Long et al., 2015; Ganin and Lempitsky, 2015; Cao et al., 2019).

Most recent methods focus on identifying classes shared by both source and target domains (common), as well as classes exclusive to the source (source private) or the target (target private) (You et al., 2019; Saito et al., 2020; Saito et al., 2020). Scenarios where only common classes are present are known as closed-set DA, while variations with only source private (partial DA), only target private (open-set DA), and both source and target private (universal DA) exist (Cao et al., 2019). For closed-set DA, techniques to align the source and target distributions through conditional adversarial training by maximizing the intra-class density have been implemented, while others using progressive adaptation of the feature norm between domains have also been proposed (Long et al., 2018; Li et al., 2021; Xu et al., 2019). For universal DA, entropy separation techniques or auxiliary adversarial discriminators have been introduced to allow the discrimination of unknown classes (Saito et al., 2020; Cao et al., 2018; Yu et al., 2023). In general, universal DA methods have been shown to perform as well or better than closed-set methods in closed DA. Nevertheless, all of these still rely on the availability of enough target data, which can become a major challenge when dealing with Few-shot scenarios (Liu et al., 2022; Zhao et al., 2021; Benigmim et al., 2023).

2.2 Few-Shot Learning

Few-shot scenarios happen when the target data available for training is very limited in comparison with the source data (Liu et al., 2022; Bashkirova et al., 2023; Motiian et al., 2017). In general, this means having 2-5 images per class. A more challenging setting, with only 1 image per class, is called One-shot (Benigmim et al., 2023; Yang et al., 2021). Overcoming both Few-shot and One-shot scenarios is a major problem for the adoption of computer vision techniques in practical settings, as these may incur in cases with limited and imbalanced data, usually due to constraints in data acquisition(Liu et al., 2022).

Few-shot learning methods have generally attempted to address this problem through meta-learning approaches. As summarized by (Zhao et al., 2021), these can be divided into three categories: rapid adaptation from source to target classes, prototypical learning through the aid of encoders, and substitution of gradient descent with novel optimization algorithms. Some of the latest Few-shot DA methods address both the Few-shot DA and Few-shot learning problems by utilizing labeled target samples to build prototypes that align with the source and target (Zhao et al., 2021). Another method (DATUM) has used diffusion models to augment data for One-shot scenarios (Benigmim et al., 2023).

2.3 Diffusion Models

Diffusion models (DMs) have represented a major step towards high-quality photo-realistic image generation, thanks to their combination with text encoders which allow for image generation through the guidance of natural language prompts (Rombach et al., 2021). Fine-grained generation has been achieved through the use of fine-tuning methods such as Dreambooth (Ruiz et al., 2022), Textual Inversion (Gal et al., 2022), and ControlNet (Zhang and Agrawala, 2023). In addition, recent advances have explored the possibility of applying these fine-tuning methods directly to the latent space of a pre-trained autoencoder, a faster alternative to standard DMs (Rombach et al., 2021). A recent work has successfully used Dreambooth to fine-tune latent DMs to address One-shot DA for image segmentation with a model-agnostic proposal (Benigmim et al., 2023). Inspired by this research, here we extend the application of fine-tuned latent DMs via Dreambooth to address multi-class classification in Few-shot DA.

3 PROPOSED METHOD

3.1 Problem Statement

We consider a set of labeled source domain images $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$, where N_s is the total number of source classes, as well as a set of unlabeled target images $D_t = \{x_i^s\}_{i=1}^{N_t}$, where N_t is the total number of target classes. The label classes for the source and target are denoted by C_s and C_t, respectively. Here, we consider the closed-set scenario without any private source and target classes, e.g., $C = C_s \cap C_t = C_s = C_t$. Let n_s^k and n_t^k be the number of images of class k, where $k \in C$.

In a Few-shot scenario, the number of images per class for the target domain is smaller than their counterparts in the source domain, e.g., $n_t^k < n_s^k, k \in C$. Other works usually consider $2 \leq n_k \leq 5$. Without losing generality, here we take $n_k = 3$. We want to generate new synthetic images for D_t using a fine-tuned latent DM model such that after data augmentation, $n_t^k \geq n_s^k, \forall k \in C$.

To simulate the Few-shot scenario using standard available data sets, we extract 3 images from each class to construct the *original Few-shot* data set and generate n additional images to build the *synthetic augmented* data set. We combine both the original 3 images with the n synthetic samples to build the total training set. The rest of the target images that were separated for the Few-shot generation are saved for testing. [1]. We found that this method of train-test separation provides a better reference to simulate a few-shot setting of a real-world scenario while using open-source data sets to benchmark model performance.

3.2 Model Fine-Tuning

Image generation through diffusion models is performed as an image-denoising task, where first an image X_0 is sequentially degraded by the addition of Gaussian noise at each forward iteration, $X_1, ...X_T$. Then, a convolutional network $\eta_\theta(.)$ is trained to reconstruct the original image from the noisy input in a backwards fashion (generally with U-Net). As this is computationally expensive, research has suggested that DMs can instead work in the latent space of a pre-trained autoencoder (Rombach et al., 2021).

Conditioning is achieved by introducing an additional input to guide the denoising process of the network. In the case of text-conditioning, the embeddings of a text encoder $\tau_\theta(y)$ are used to augment the embedding of the network $\eta_\theta(X, y)$ using a cross-attention mechanism. This modifies the loss function of the DM as described in (Ruiz et al., 2022).

For Dreambooth (Ruiz et al., 2022), a prompt with a unique identifier (e.g. *zwx*) is assigned to the input images to fine-tune the DM weights (e.g. p="a photo of zwx backpack"). Thus, the model retains the knowledge already present in the pre-trained weights (e.g. general appearance of a backpack), while still capturing the specific target domain features of the fine-tuning samples (e.g. type of the backpack present in target domain). In particular, we use the implementation of Dreambooth with StableDiffusion 1.4 by Shivam Shirao as provided in (Shirao, 2022).

Initially, we considered two fine-tuning methods to generate the augmented synthetic data:

- Single Model: we fine-tune a single model for each class using all 3 images, with a single prompt with and a unique identifier. So, we have one fine-tuned DM model for each class.
- Triple Model: we fine tune one model for each one of the 3 images, using the same prompt for each case but different unique identifiers. So, we have 3 fine-tuned DM models for each class.

The single model approach has the advantage of being faster and easier to implement, as only n_c models are required for the classification task (n_c is the number of classes), but it may be sensitive to low-quality image generation or lead to generation of bad samples for DA. The triple model approach is still easy to implement but requires $3n_c$ models for the classification task; this could help in generating more high-fidelity images, but at the expense of a potential lack of diversity. Detailed studies (Section 5) showed that the single model method performs similarly to the triple model method in most cases, so we opted for the former when comparing the performance of our method with different domain adaptation methods.

[1] For clarity: (1) Training phase: the DA model learns from all the source data, the few-shot target data, and the synthetic augmented target data. (2) Testing phase: the DA model tests its results on the remaining target data only (the one never seen during training)

3.3 Synthetic Augmentation

3.3.1 Image Generation

After fine-tuning each DM model using Dreambooth, we inputted prompts with their unique identifiers to generate new target synthetic images and analyzed the quality of image generation. Using simple prompts (e.g. p="photo of zwx backpack") generated images that either lacked diversity or were likely to be combined with other objects and/or background elements. To address this, we repeated the process using targeted prompts that captured the image context (e.g. p="photo of zwx backpack leaning against wall"). The generated results had more diversity and were more representative of the target domain for that class. We also analyzed the effect of the number of training steps, inference steps, and guidance scale, finding that the optimal values for all classes were: $training\_steps$: 400, $inference\_steps$: 100, $guidance\_scale$: $6.5 - 7.5$. This applied both to the single model and the triple model. We repeated this process for each of the classes in the target domain and constructed the full target data sets by combining the original limited data (few-shot) with the synthetic (augmented) data sets.

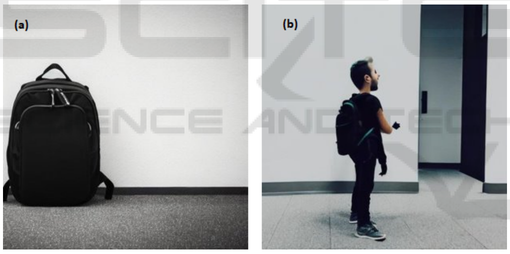


Figure 1: Good synthetic sample (a) versus bad synthetic sample (b) generated with Dreambooth for Office31-Webcam class 'backpack', using the same prompt, ("photo of zwx backpack against wall", guidance scale=6.5).

3.3.2 Manual Selection

Despite our best efforts at model fine-tuning, we noted that some of the generated samples could produce negative transfer[2] due to irrelevant objects appearing on the image (e.g., a person with a backpack), or objects with anomalous shapes (e.g., a scissor with more than two blades) (Figure 1). Automation was attempted using a combination of several metrics calculated between original and synthetic images (e.g., FSIM, RMSE, SSIM, UIQ) to filter out *bad samples*. However, no effective combination that worked for all classes was found. To determine if these *bad samples* would really incur in degraded model performance, we manually extracted a set of *good samples* based on the following criteria: (i) the object must be alone on the image (e.g., no presence of irrelevant objects), (ii) the object must represent the target domain (e.g., same background), (iii) the object must not have any visible aberrations or deformations (e.g., backpack with multiple handles).

We performed tests for each model using both the manually selected samples (chosen) and all the unfiltered synthetic data (all).

3.4 Domain Adaptation

As the data augmentation step is independent of DA, this approach can be implemented with any model, with the only minor change required being the adaptation of the testing procedure (Section 3.1). We considered the following four universal DA models and two closed-set DA models:

- DANCE. Domain Adaptative Neighborhood Clustering via Entropy optimization (DANCE) relies on learning well-clustered features from source and target to identify common classes at the mini-batch level, followed by entropy-based alignment to discriminate target samples as common classes target private classes (Saito et al., 2020).
- (NUDA). Noisy-Universal Domain Adaptation uses two classifiers with different initializations trained in an adversarial manner with a generator to filter out samples with noisy labels, and the distributions of the source and target domain are aligned by minimizing the divergence between the outputs of the classifiers (Yu et al., 2023). The model is robust even for non-noisy cases, so we consider it despite not having any noisy labels.
- ETN. Example Transfer Networks (ETN) learn domain-invariant representations across the source and target domains via a progressive weighting scheme using an auxiliary domain discriminator, which quantifies the degree of domain *transferability* of source samples while at the same time controlling the importance of learning in the target domain (Cao et al., 2019).
- UAN. Universal Adaptation Networks (UAN) attempt to learn domain-invariant representations across domains by quantifying the sample-level *transferability* to automatically identify common samples between source and target and filter out unknown samples (You et al., 2019).

[2] Negative transfer: when the accuracy of a model after domain adaptation is worse than the model accuracy of the model without domain adaptation.

Table 1: Mean model accuracies for each universal domain adaptation model per number of classes (part 1). (a) refers to using all synthetic augmented data, while (b) refers to using synthetic data with manual selection.

Classes	SO	DANCE	DANCE+DB (a)	DANCE+DB (b)	NUDA	NUDA+DB (a)	NUDA+DB (b)
31	0.80	0.82	0.75	0.79	0.84	**0.87**	0.85
20	0.81	0.82	0.79	0.84	0.89	**0.94**	0.88
10	0.88	0.91	0.88	0.86	**0.95**	0.94	**0.95**

Table 2: Mean model accuracies for each universal domain adaptation model per number of classes (part 2). (a) refers to using all synthetic augmented data, while (b) refers to using synthetic data with manual selection.

Classes	SO	ETN	ETN+DB (a)	ETN+DB (b)	UAN	UAN+DB (a)	UAN+DB (b)
31	0.80	**0.87**	**0.87**	**0.87**	0.35	0.36	0.36
20	0.81	**0.93**	**0.93**	**0.93**	0.62	0.62	0.62
10	0.88	0.97	0.97	**0.98**	0.97	0.97	0.97

- CDAN. Conditional Adversarial Domain Adaptation (CDAN) captures cross-variance between feature representations and classifier predictions (multi-linear conditioning) and prioritizes the discriminator on samples by using entropy-aware weights (Long et al., 2018).
- MMD. Maximum Density Divergence (MMD) is a distance loss that quantifies the distribution divergence, which, coupled with standard adversarial loss, is used to minimize inter-domain divergence and maximize intra-class density to align and compact class distributions (Li et al., 2021).

4 COMPUTER EXPERIMENTS

4.1 Experimental Set-Up

We consider a closed-set scenario using the Office31 data set for multi-class classification. We simulate the Few-shot scenario by picking 3 representative samples for each target class: representative images must have different poses of the object and/or display different variations of the class. We justify this selection by arguing that in real-world scenarios, limited samples will try to capture at least the most representative cases of each of the target domain classes. We then fine-tune a single DM model for each class and use targeted prompts to generate 32 additional synthetic images for each class. We considered 400 training steps, learning rate of 1e-6, with prior preservation weight of 1.0. For manual selection, we chose 10 images for each class, following Section 3.3.2.

We randomly pick 20 (Office20) and 10 (Office10) sets of classes from Office31 to compare the performance of the models as a function of class number. For baseline, we use ResNet50 trained only on source data, as previous works have shown it outperforms most standard models in image classification (Liu et al., 2022). Model-only case denotes the performance of the base model with Few-shot target data (3 samples for each class), Model+Dreambooth (a) denotes its performance with all the synthetic data augmented target samples (3+32 samples for each class), and Model+Dreambooth (b) denotes its performance manually chosen synthetic data augmented target samples (3+10 samples for each class). All models were trained for 5000 steps with a batch size of 32, and results are reported at the end of the training and rounded up to two decimal places.

Hyperparameter tuning was previously performed for each model, and the results for the best accuracies of each model were considered. For DANCE and NUDA, a margin of 0.5 and a threshold of $log(N_c)/2$ were used, with N_c being the number of classes. For NUDA, the function to artificially introduce noisy labels was removed from the original code. For ETN, weights for adversarial augmented loss trade-off and adversarial loss trade-off were 10.0 and 1.0, respectively. For UDA, the weight for transferability trade-off was -0.5. For CDAN and MDD, the same configurations as their original works were utilized (adapting for the number of classes).

4.2 Results

Results for the mean accuracies for each of the models and cases considered are shown in Tables 1 and 2.

For NUDA, mean accuracy improves between 1% and 3% when using augmented data in comparison with model-only. In particular, not applying manual selection (a) and instead considering all the generated images (b) yields better results. This is likely due to the noisy label detection module as well as the entropy separation from the target, which is able

Table 3: Mean model accuracies for each closed-set adaptation model per number of classes. (a) refers to using all synthetic augmented data, while (b) refers to using synthetic data with manual selection.

Classes	SO	CADN	CADN+DB (a)	CADN+DB (b)	MDD	MDD+DB (a)	MDD+DB (b)
31	0.80	0.83	0.85	**0.87**	0.84	0.84	0.85
20	0.81	0.90	**0.93**	0.90	0.92	0.89	0.89
10	0.88	0.92	0.90	0.87	**0.98**	0.94	0.94

Table 4: Mean per-class accuracy for each model for the 10 class scenario. (a) refers to using all synthetic augmented data, while (b) refers to using synthetic data with manual selection. We consider the best cases for each method. For NUDA, per-class accuracy for the first classifier is reported.

Class	SO	DANCE+DB (b)	NUDA+DB (a)	ETN+DB (b)	UAN+DB (a)	CDAN+DB (b)	MMD (a)
backpack	0.90	0.92	1.0	1.0	0.96	**0.97**	1.0
bike	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
calculator	0.47	0.5	0.82	0.82	**1.0**	0.43	0.71
headphones	0.82	0.96	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
keyboard	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
laptop	0.67	0.90	0.87	**1.0**	0.89	0.87	0.93
monitor	0.85	0.91	0.95	0.97	**1.0**	0.74	0.93
mouse	**1.0**	**1.0**	**1.0**	**1.0**	0.98	0.97	0.97
mug	0.99	0.95	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
projector	0.95	0.73	0.77	**0.97**	0.91	0.83	0.93

to sort the bad synthetic samples that could produce negative transfer. For ETN and UAN, there is no significant change in overall accuracies, so while the augmented data does not necessarily improve model performance, it does not hinder it either. The results are similar for all augmented data and manually selected data. This is likely thanks to the auxiliary discriminators in these models, which are already filtering/re-weighting potential bad synthetic samples. For DANCE, model performance is slightly worse for synthetic augmented data, while results for manually chosen data are comparable to the model-only case. This could indicate the presence of negative transfer from bad synthetic samples, as this model does not have a module to filter out/weight out samples that may negatively affect the DA. This could also be caused by the neighbor clustering module, which be produce an unstable classification border for the classes.

CDAN shows an increase in mean accuracy of up to 4% for 31 classes when considering manual selection, while it has an increase of 2% when using all synthetic data. In contrast, in the the other cases, there is no significant improvement. This is likely because of its multi-lineal conditioning, which may be sensitive to the distribution of synthetic data in the latent space when dealing with a lower number of classes but has a positive impact as more classes are considered. MMD does not display any increase in accuracy considering augmented data, and there is even a decrease in accuracy for 20 and 10 classes when compared with the model trained only on Few-shot data. This could be explained by the distribution of synthetic data, which may play against the minimization of the inter-class density of this method, in addition to the fact that there is no technique to filter out bad samples that may not closely align to the distribution of the original data.

Table 3 presents the results of the mean per-class accuracies for each model for the 10 class Few-shot scenario. We can observe the positive effect of synthetic augmented data at the per-class level, as almost all classes have their classification accuracy improved in the Few-shot scenario. In general, most of them have an average increase of 3% to 10%, with the exception of the class ”calculator”, which is able to improve by up to 30% when using NUDA and ETN. Some classes in DANCE and CDAN present worse accuracy (e.g., ”projector”), probably due to negative transfer, but no significant negative transfer is seen when using NUDA and ETN (sauf for ”projector”), so these methods seem to be robust when combined with our approach.

Table 5: Mean per-class accuracy for DANCE-only, DANCE+Dreambooth with augmented data using a single-model (I), and DANCE+Dreambooth with augmented data using a triple-model (II).

Class	DANCE-only	DANCE+DB(I)	DANCE+DB(II)
backpack	0.71	0.83	**0.97**
bike	**1.0**	**1.0**	**1.0**
calculator	**0.78**	0.43	0.50
headphones	0.82	0.94	**0.99**
keyboard	0.80	**1.0**	**1.0**
laptop computer	0.61	0.75	**0.95**
monitor	0.74	0.77	**0.86**
mouse	0.90	**0.93**	0.92
mug	0.85	0.89	**0.92**
projector	0.68	**0.91**	0.70

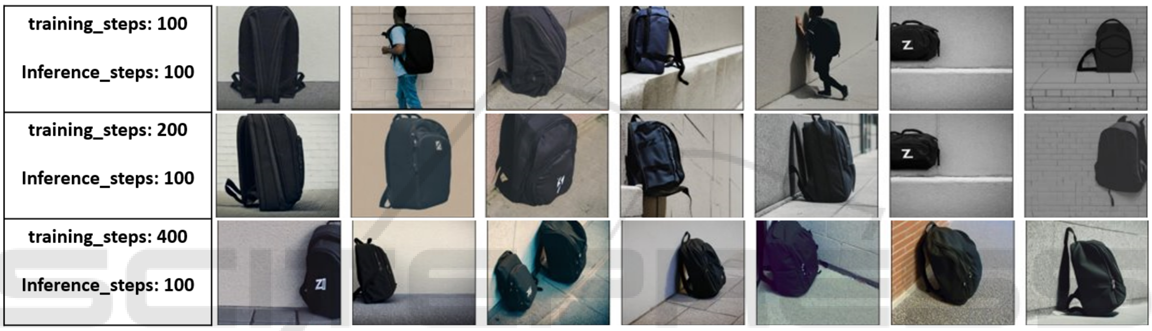


Figure 2: Effect of training steps in DM fine-tuning for Dreambooth image generation.

5 ABLATION ANALYSIS

5.1 Hyperparameter Tuning

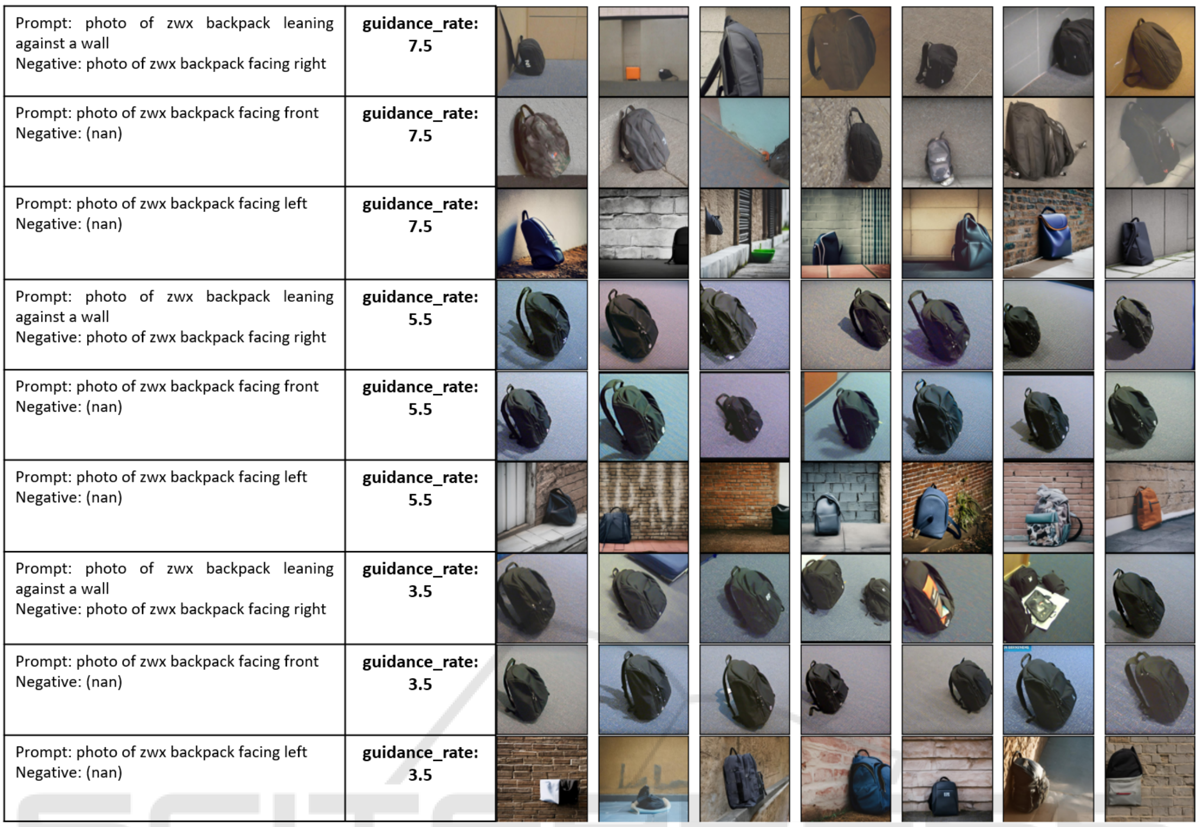
We studied the effect of Dreambooth hyperparameters on the quality of image generation with the fine-tuned diffusion model. First, we considered the number of training steps and generated images with 100 inference steps using the same prompt. Increasing the number of training steps generally leads to a better representation of the target domain and visualization of the object in the generated prompts (Figure 2).

We also explored the effect of the guidance rate on the quality of image generation for fixed training steps. Results for the class *backpack* are shown in Figure 3. Both for the single-model and the triple-model, the optimal values were in the range of 6.5-7.5, usually with the higher end being better for objects with more complex compositions. On the other hand, more specific prompts (e.g., p=”photo of zwx backpack leaning against wall facing left”) did not affect image diversity in comparison with already well-targeted prompts (e.g., p=”photo of zwx backpack leaning against wall”), so we opted for the later one. Negative prompts did not have a significant effect on image quality generation.

5.2 Single-Model versus Triple-Model

Considering the DANCE algorithm as a backbone for DA, we studied the effect of using the single-model (3 original images, 24 synthetic images) versus the triple-model approach (11+1 original images, 8+8+8 synthetic images) on model accuracies. The results in Table 4 differ slightly from the ones in Table 3 since this analysis was performed with a smaller set of augmented synthetic data (10 images for each case) and calculating mean accuracy for the last 2000 iterations.

While the use of synthetic data does improve mean performance with respect to the non-augmented data case, there is not much difference between overall accuracies with the single-model (I) and triple-model (II). Some classes like *laptop computer* and *monitor* show an improvement for the triple-model. In contrast, the class ”projector” shows a decrease in



Prompt: photo of zwx backpack leaning against a wall Negative: photo of zwx backpack facing right	**guidance_rate: 7.5**
Prompt: photo of zwx backpack facing front Negative: (nan)	**guidance_rate: 7.5**
Prompt: photo of zwx backpack facing left Negative: (nan)	**guidance_rate: 7.5**
Prompt: photo of zwx backpack leaning against a wall Negative: photo of zwx backpack facing right	**guidance_rate: 5.5**
Prompt: photo of zwx backpack facing front Negative: (nan)	**guidance_rate: 5.5**
Prompt: photo of zwx backpack facing left Negative: (nan)	**guidance_rate: 5.5**
Prompt: photo of zwx backpack leaning against a wall Negative: photo of zwx backpack facing right	**guidance_rate: 3.5**
Prompt: photo of zwx backpack facing front Negative: (nan)	**guidance_rate: 3.5**
Prompt: photo of zwx backpack facing left Negative: (nan)	**guidance_rate: 3.5**

Figure 3: Effect of specific prompts and guidance scale in DM fine-tuning for Dreambooth image generation.

accuracy, as the diversity from the original data set was probably not captured by the triple-model.

Therefore, while it seems that the triple-model yields slightly better results at the per-class level, with small improvements in mean accuracy, it becomes more computationally expensive when considering a greater number of classes. For this reason, we opted for the single-model approach for our experiments with multiple DA methods.

5.3 Effect of Image Selection

While DANCE and CDAN have a clear improvement when considering manual image selection, this selection becomes unnecessary for methods that are able to filter out bad samples, such as Non-Noisy UDA and ETN (Section 4). So, while manual selection may be convenient for scenarios with a small class number, in cases with time or budgetary constraints, it is not strictly necessary to perform additional selection as long as a proper DA method is implemented.

5.4 Effect of Number of Classes

The best results for DA are generally obtained for the 10-class case, which is expected as models are less likely to misclassify objects in this case (Section 4). In addition, the fewer classes considered in Few-shot scenarios, the less likely it is that bad samples from synthetic image generation will affect model performance. However, the greatest improvements when comparing the model with synthetic augmented data with respect to the ones without augmented data can be observed for the 31-class case, in particular for NUDA and CDAN. This strongly suggests that number of classes, something that may be common in industries that deal with a wide variety of different products and objects.

6 CONCLUSIONS

We have presented an innovative approach using fine-tuned diffusion models with Dreambooth for synthetic data augmentation to address the problem of Few-shot multi-class classification in domain adaptation. We have detailed the steps and considerations to follow when performing fine-tuning for diffusion model-assisted data augmentation and how to combine it with state-of-the-art DA models. We observe that we can generate additional synthetic data that captures the target domain for each class and improves model accuracies over their non-augmented counterparts. Our approach is model-agnostic and easy-to-implement, converting the Few-shot problem into a standard problem of DA. While not all DA models (e.g., UAN) may benefit from this approach, other methods do show an improvement in their average and per-class accuracy (e.g., NUDA, CDAN), showcasing the prospective application of this technique to real-world scenarios. This is the first work that, to our knowledge, has considered this combination of these methods to address multi-class classification.

Future work could consider the open-set and partial open-set cases for few-shot scenarios to study how the presence of potentially bad synthetic samples may affect the accuracy in the presence of unknown classes, as well as a more exhaustive study on the trade-off between the single-model and triple-model fine tuning strategies.

ACKNOWLEDGEMENTS

The following research was financed with a fellowship from the Japan International Cooperation Agency (JICA), as part of its Japan-Mexico Training Program for the Strategic Global Partnership 2023. We would also like to thank the Graduate School of Information and Engineering at Ritsumeikan University Campus for its kind support.

REFERENCES

Bashkirova, D., Mishra, S., Lteif, D., Teterwak, P., Kim, D., Alladkani, F., Akl, J., Calli, B., Bargal, S. A., Saenko, K., Kim, D., Seo, M., Jeon, Y., Choi, D.-G., Ettedgui, S., Giryes, R., Abu-Hussein, S., Xie, B., and Li, S. (2023). Visda 2022 challenge: Domain adaptation for industrial waste sorting.

Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V., and Lathuilière, S. (2023). One-shot unsupervised domain adaptation with personalized diffusion models. *arXiv preprint arXiv:2303.18080*.

Cao, Z., Ma, L., Long, M., and Wang, J. (2018). Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Cao, Z., You, K., Long, M., Wang, J., and Yang, Q. (2019). Learning to transfer examples for partial domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2980–2989.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. 17(1):2096–2030.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., and Shen, H. T. (2021). Maximum density divergence for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3918–3930.

Liu, X., Yoo, C., Xing, F., Oh, H., Fakhri, G., Kang, J.-W., and Woo, J. (2022). Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France. PMLR.

Long, M., CAO, Z., Wang, J., and Jordan, M. I. (2018). Conditional adversarial domain adaptation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Motiian, S., Jones, Q., Iranmanesh, S. M., and Doretto, G. (2017). Few-shot adversarial domain adaptation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6673–6683, Red Hook, NY, USA. Curran Associates Inc.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2022). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

Saito, K., Kim, D., Sclaroff, S., and Saenko, K. (2020). Universal domain adaptation through self supervision. In

Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16282–16292. Curran Associates, Inc.

Shirao, S. (2022). Github: Stable diffusion + dreambooth. https://github.com/ShivamShrirao/diffusers/tree/main /examples/dreambooth.

Xu, R., Li, G., Yang, J., and Lin, L. (2019). Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1426–1435.

Yang, C., Shen, Y., Zhang, Z., Xu, Y., Zhu, J., Wu, Z., and Zhou, B. (2021). One-shot generative domain adaptation.

You, K., Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2019). Universal domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, Q., Hashimoto, A., and Ushiku, Y. (2023). Noisy universal domain adaptation via divergence optimization for visual recognition.

Zhang, L. and Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs].

Zhao, A., Ding, M., Lu, Z., Xiang, T., Niu, Y., Guan, J., and Wen, J.-R. (2021). Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1390–1399.