SYNTHETIC-TO-REAL UNSUPERVISED DOMAIN ADAPTATION FOR IMAGE AND VIDEO CLASSIFICATION

by
Arushi Sinha

An independent study report

Baltimore, Maryland
December, 2023

Abstract

Deep Learning solutions for activity recognition across viewpoints need a large amount of data for training. Collection and labeling of these large datasets can be time consuming, expensive and noisy. Thus, synthetic data generated using graphics or large-scale generative models, will be useful for data augmentation. For a model trained on this synthetic source domain to generalize well on real target domain which may have variation in viewpoints, one needs to mitigate or compensate for domain shift.

In this study, we aim to translate the synthetic images and videos generated from rendering engines to appear realistic by leveraging denoising diffusion methods and large language models, so that activity recognition from ground to aerial views can be accomplished.

Research Advisor

Dr. Rama Chellappa (Primary Advisor)
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Contents

List of Figures

Chapter 1

Introduction

Problem Statement

In the field of machine learning, data domain variation presents a significant challenge for achieving optimal performance across various data domains. This issue arises primarily due to the inherent complexities and costs associated with collecting real-world data along with its corresponding ground truth labels or annotations. The arduous nature of this process often necessitates the use of simulated or synthetic data generated by graphics or large-scale deep neural networks.

The reliance on these alternative data sources introduces a substantial disparity in data distribution when compared to real-world data. For predictive models to effectively generalize to target domains that differ from their initial training domains, innovative solutions must be explored to address this domain variation. For instance, models trained entirely on synthetic data frequently exhibit poor performance when applied to real-world scenarios. One potential strategy involves training models on a combination of both synthetic and real-world data. However, this approach does not always achieve the same level of accuracy as models trained solely on real data.

The core challenge, therefore, lies in enabling models trained on synthetic or a mix of synthetic and real data to match the predictive or recognition capabilities of models trained exclusively on real-world data. This task can be approached through the implementation of Domain Adaptation algorithms. These algorithms aim to bridge the gap between varied data domains, thus enhancing the model's ability to function effectively across different types of data. Addressing this challenge is crucial for advancing the field of image recognition and ensuring the applicability and reliability of these models in real-world settings.

To address this issue, our research aims to translate synthetic images and videos, generated from rendering engines, to appear more realistic. We adapt pre-trained diffusion models (such as Stable Diffusion) trained on large amounts of real-world data for image and video translation. We also explore the use of image captioning models for inducing realism. This approach is designed to enhance the realism of synthetic data, thereby bridging the gap between synthetic and real-world data domains.

Datasets

For experiments, image and video datasets are selected with both synthetic and real domains.

The following image datasets are used:

VisDA-2017: VisDA-2017 is a domain adaptation dataset focused on simulation-to-real scenarios, with 280,000 images in 12 classes across training, validation, and testing domains. Training images are rendered from various perspectives of the same object, whereas validation images are sourced from the MSCOCO dataset. This dataset is specifically designed to test and improve models in adapting from simulated to real-world image contexts. The images can be observed in fig 1-1.

S2RDA-49: The S2RDA-49 dataset, essential for visual domain adaptation research, consists of 49 classes with 588,000 synthetic source domain images from ShapeNet and 60,535 target domain images from diverse real-world sources including ImageNet, ObjectNet, VisDA-2017, and the web. Compared to VisDA-2017, it has a more complicated target domain and more realistic synthetic source domain. The images

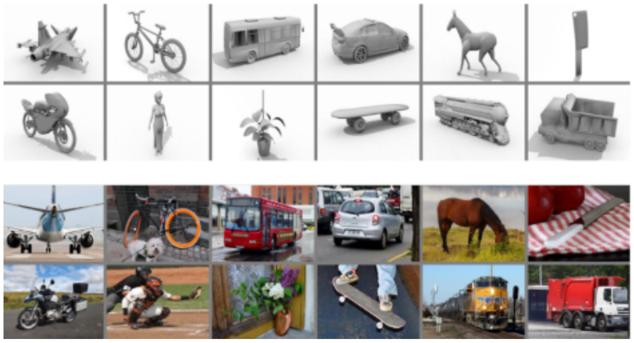


Figure 1-1. Synthetic source domain images (top) and real target domain images ((bottom) of VisDA-2017 dataset

can be observed in fig 1-2

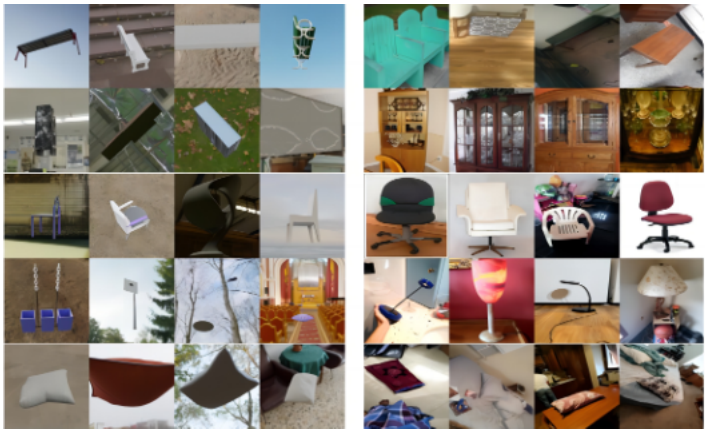


Figure 1-2. Synthetic domain images (left) and the images from real domains (right) of S2RDA-49 dataset

The following video dataset is used:

RoCoG-v2: RoCoG-v2 (Robot Control Gestures), is a dataset used for video domain adaptation, both from synthetic-to-real and ground-to-air perspectives. It comprises over 100,000 synthetic videos depicting human avatars executing gestures across seven classes. Additionally, it includes real human videos performing identical gestures, captured from both ground and aerial viewpoints. The video frames can be seen in fig 1-3.

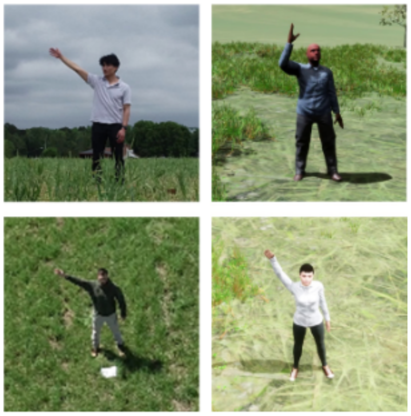


Figure 1-3. Real video frames (left) and simulated video frames (right) of RoCoG-v2 dataset,.

Related Work

Style Transfer

Neural style transfer is an image-to-image translation method, introduced by [1]. In this method, a new image is generated from either white noise or an image from which we extract content representation. Two representations are learnt using a Deep Convolutional Neural Network: (i) Content representation from the content image and (ii) Style representation from the style image. Gram matrices of the neural activations from different layers of a CNN are used to represent the artistic style of an image. Then it used an iterative optimization method to generate a new image from white noise by matching the neural activations with the content image and the Gram matrices with the style image. Now, this task can also be seen as a distribution alignment task, which is essentially the core in domain adaptation. Style transfer by matching Gram matrices is theoretically equivalent to minimizing the Maximum Mean Discrepancy [2].

Domain Adaptation via Style Transfer

In this [3] work, domain adapatation is implemented by taking advantage of style transfer and adversarial training to predict pixel perfect depth from a single real-world color image based on training over a large dataset of synthetic environment. The real-world input RGB images are stylized to appear synthetically generated using a CycleGAN. The monocular depth estimation model is then trained on these images to perform well on test real-world RGB images. This work proved to perform better than contemporary state-of-the-art methods of depth estimation. Hence, a GAN based style transfer approach to adapt real-world data to fit into the distribution which is used to train a model for predictive/estimation task to perform better on real test data seems promising.

In another work [4], realism of synthetic images is improved using unlabelled real data to ensure that the model does not overfit to 'unrealistic' details in synthetic data. The synthetic images are refined using adversarial training using a refiner and discriminator network. The discriminator ensures that the generated refined images are indistinguishable from real ones and the refiner generates the synthetic images as well as penalizes large changes between synthetic and refined images. Training a CNN on these refined images outperforms state-of-the-art gaze estimation models. Even for another task of hand pose estimation from depth images, training a model with output of SimGAN performs relatively better than the ones trained on real images with supervision.

Diffusion Models

Diffusion models originated in 2015 to learn a model that can sample from highly complex probability distribution. They have emerged as a transformative force in generative modeling, with pioneering work by Sohl-Dickstein et al. [5] introducing the concept of gradually transforming noise into structured data. This technique, fundamentally different from the methods used in Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), involves a forward trajectory of adding noise and a reverse trajectory where a neural network is trained to periodically denoise. Ho et al. [6] demonstrated the remarkable capabilities of diffusion models in generating high-fidelity, detailed images. Their work showed proficiency in capturing complex style patterns, leading to increasing applications in image synthesis and editing. One notable application of diffusion model is text-to-image generation. Various models like Stable Diffusion [7], Imagen [8], and GLIDE [9] have shown promising results in high quality image generation that leads to support downstream tasks from these resulting images.

Now, for some task specific use-cases, the need to have more control over this generation process increases. The work by Zhag et al. [10] introduces ControlNet, a class of models that provide more control to the user by conditioning Stable Diffusion in addition to the text prompt control. It does so by preserving their core capabilities through locked parameters, while introducing a trainable copy of encoding layers connected via zero-initialized convolution layers, ensuring stable training without introducing harmful noise to the pretrained backbone. The work of Gal et al. [11] introduces another way of adding more control to the generation of these large, pre-trained models. This method learns to represent the concept of style or objects from a few images through new words in a the frozen text-to-image model for personalized image generation. Essentially, textual Inversion creates an additional "word" that is added to the base model's vocabulary so it can generate it by giving it a new point to try to get close to as it removes noise. Ruiz et al. [12] fine-tune the diffusion model directly to learn the representation of the subject in images. This personalized generation has shown potential to improve downstream tasks such as image classification, editing, and segmentation. For the improvement in the classification task, see the experiment section.

Image Style Transfer Using Diffusion Models

In a setting where we have a large unlabelled target dataset, the traditional UDA methods reduce the domain gap by learning a model using both labelled source data and unlabelled target data for it to work well on the target domain. But these methods may not estimate and align distributions well in a setting where only a single unlabelled datum is available. This problem is called one-shot-unsupervised domain adaptation (OSUDA). These methods use the approach of overpopulating the target dataset by transferring the style of the target domain to the source data. Now, since information about the target domain is insufficient as only a single datum is available, and there is an issue where the data's scene content and spatial layout is constrained to what the source data can offer. To overcome these issues, the work [13] propose their method of augmenting the target dataset by leveraging denoising diffusion methods, so that generated images go beyond simply imitating the target domain and instead incorporate a range of different and realistic scene arrangements. This is achieved by leveraging the abundant prior knowledge encoded in the DMs.

Video Style Transfer

In videos, the domain adaptation task becomes more extensive since the source and target domain not only differ spatially, but also temporally. Images are spatially well-structured data, while videos are sequences of images with both spatial and temporal relations [14]. Efforts to implement style transfer in videos cannot be applied considered without considering temporal consistency.

Synthetic data

The use of synthetic data has increased since the recent advancements in text-to-image and text-to-video models. It is especially helpful in vision tasks that require extensive annotations. Azizi et al. [15] show that the augmenting the training dataset with samples from the powerful, fine-tuned large text-to-image diffusion models results in state-of-the-art FID and classfication accuracy scores compared to ResNet and Vision Transformer baselines. The work by He et al. investigate the use of this synthetic data from state-of-the-art generative models image recognition. The results show that they are better suited in zero-shot and few-shot recognition as they perform significantly better. But, in cases when the model is initially trained using only synthetic data, it typically fails to achieve satisfactory performance, showing lower data efficiency and effectiveness in handling classification tasks compared to models trained with real data. Other drawbacks of synthetic data are also mentioned in the work by Lu et al. [16]. They raise the issues of data privacy and fairness as the objective of data synthesis to maintain the distribution of the original data may result in synthesized data inheriting these attributes. Generating synthetic data that mirrors real-world statistical properties can also inadvertently inherit biases from data preprocessing, collection, and algorithmic processes. These potential biases could be a significant concern.

Chapter 2

Method

Models

Image Translation Model

Our image generation model is built using the Stable Diffusion (SD) v1.5 model. We fine-tune this model using Dreambooth, and modify its embedding space using Textual Inversion. Additionally, ControlNet is used to condition this trained, *personalized* SD v1.5 model.

1. **Architecture and Training Details:** Stable Diffusion v1.5 is trained on 512x512 images from a subset of the LAION-5B database [17]. It is based on the Latent Diffusion model. The model specification is described in table 2-I. Two libraries: diffusers [18] and transformers [19] are utilized to get the base model's checkpoints.

   For training this base model, Dreambooth is applied using the target set images of the dataset, to learn the style of the target domain. During training, we applied prior preservation loss, leveraging the model's self-generated samples to enhance its capacity for diverse image generation. Prior preservation loss was set to 1.0 along with class prompt *"an object"*. We employed a resolution of 512 for high-quality output, a single instance per batch for intensive learning, and used 8-bit Adam optimizer alongside gradient checkpointing for efficiency. The training involved a learning rate of 2e-6, without any warmup steps, over 800 steps with 1000 class images. The instance prompt used for this training: *"an object in the style of sks object"*, where *sks* is the identifier used that captures the style of target domain. This training takes 2 hours using one GeForce RTX 2080 GPU.

For more conditioned generation, a new word that represents the style of target set images is added to the trained SD v1-5 model using Textual Inversion. This training focused on the learnable property "object," using the placeholder token *<realism>* and initializer token *"valid"* to guide the model's learning towards a specific concept. Key parameters of this training included a high-resolution setting of 512 and a single instance per batch, coupled with a gradient accumulation of four steps for effective learning optimization. The model was trained for 3000 steps, with a learning rate of 5.0e-04, scaled appropriately and maintained constant throughout the training without any warmup steps. The training time was 1 hour using one GeForce RTX 2080 GPU.

Version	1.5
VAE	AutoencoderKL
UNet	UNet2DConditionModel
Text Encoder	CLIP
Tokenizer	CLIPTokenizer
Inference Steps	20
Scheduler	PNDMScheduler
Guidance Scale	9

Table 2-I. Model Specification of SD v1-5

2. **Image Translation:** This trained model is then integrated with ControlNet for the image-to-image translation task. This StableDiffusionControlNet pipeline utilizes the canny edge map version to help retain the structure of objects during translation. The prompts used are in the form of *"a class_name in the style of valid, best quality, extremely detailed, in the style of sks object"*. The class_name

addition to the pormpt is to ensure class specific translation. In addition to this, a default negative prompt is also added in the translation process: *"longbody, lowres, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality"*. The specifications of the trained SD v1.5 model is unchanged and can be seen in table 2-I.

Classification Model

The classification model used for evaluation is a Swin Transformer backbone, pre-trained on ImageNet dataset.

1. **Architecture:** To preserve the learned features and transfer the knowledge from the pre-trained swin model, we froze its weights during training and the classification head is modified according to the dataset used. The architecture details are below:

Head Architecture: Customized Classification Head	
Component	**Specification**
Initial Layer	Linear layer reducing to 512 dimensions
Activation Function	ReLU (Rectified Linear Unit)
Regularization	Dropout (30% rate)
Final Layer	Linear layer mapping to target class count
Optimizer	AdamW
Loss Function	Label Smoothing Cross-Entropy

Table 2-II. Classification Head Architecture Specifications

2. **Training Details:** In our training setup, the hyperparameters were carefully chosen to optimize the model's performance. The training was conducted over 7 epochs. Batch size for the training set was set to 128, and for the validation and test set it was set to 32. For the learning rate, a dual strategy was employed: the backbone of the model, being pre-trained, was assigned a low learning rate to preserve its learned features, while the newly added head of the model had

a higher learning rate, ranging from 0.0001 to 0.001, to facilitate more rapid learning of the new task-specific features. Additionally, we used the StepLR learning rate scheduler with a step size of 3 and a gamma value of 0.97, allowing for a gradual decrease in the learning rate as the training progressed, thereby ensuring more stable and effective model optimization.

Algorithm

- Fine-tune SD v1.5 with unlabelled target domain images as reference to capture the style information.
- Modify the text embeddings of this model by including a new embedding (tied to a special word used in the prompt), that captures the style of the target domain in its text encoder.
- Integrate ControlNet model to condition the generation of fine-tuned SD v1.5 model.
- Translate synthetic source images using this customized pipeline by passing the class name as argument for class specific translation.

A visual explanation of the method can be seen in fig 2-1

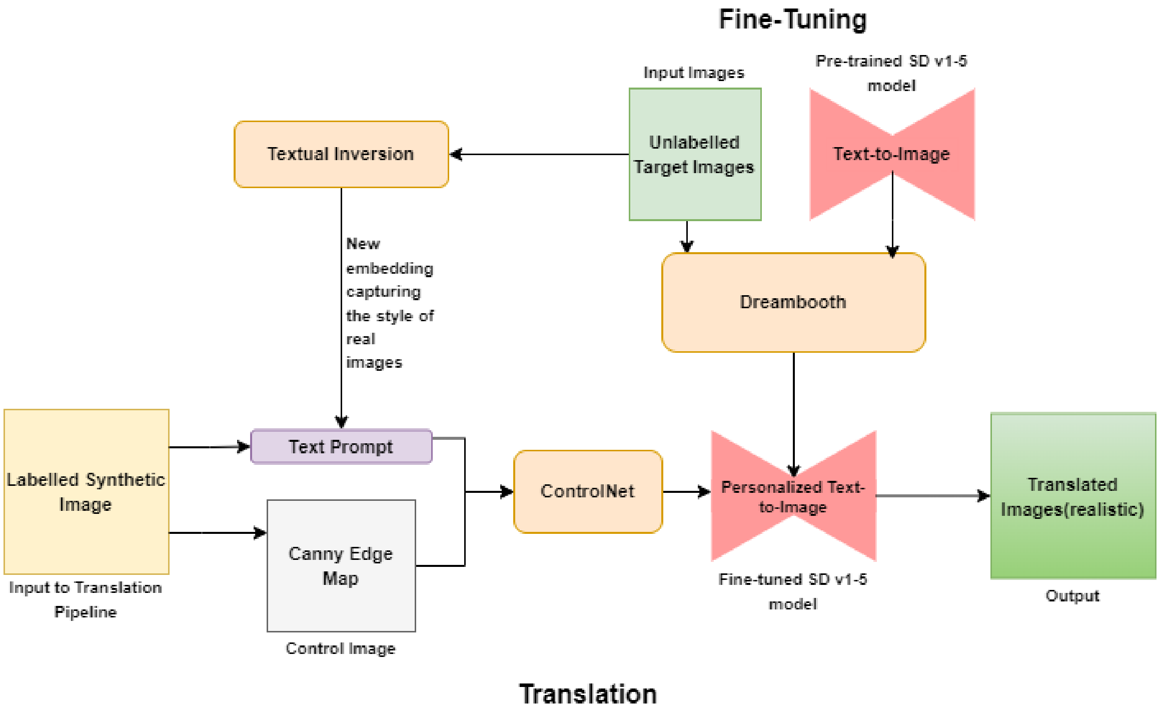


Figure 2-1. Visual Explanation of Method

Chapter 3

Experiments

Experiments for image translation and classification using two datasets are mentioned in this chapter.

Data

To evaluate the method mentioned in previous chapter, experiments are conducted on two datasets: VisDA-2017 and RoCoG-v2. The details of these datasets are provided in chapter 1.

- VisDA-2017: All the images of twelve classes (aeroplane, bicycle, bus, car, horse, knife, motorcycle, person, plant, skateboard, train, and truck) are used for translation and classification.
- RoCoG-v2: One class ("advance") is used for translation experiments.

Implementation Details

Image Translation

1. **VisDA-2017:** Stable Diffusion v1.5 model is fine-tuned using all the target domain images (validation set). Dreambooth is used which generated a new personalized model, and a newly learnt embedding of the target style. For

training this SD v1.5 model, 55,388 target images are used. These images are also used to generate a new embedding represented by a token ("valid") using Textual Inversion. The personalized model along with new learnt embedding are then used to perform translation. All the synthetic images of each class are translated using this new generative model conditioned with ControlNet. For each synthetic image, its canny version is generated and passed into this model along with the prompt *"a class_name in the style of valid, best quality, extremely detailed, in the style of sks object"*. This prompt contains the new word "valid" which maps to the new embedding and an identifier "sks" learnt while training the model. This experiment leads to a new dataset of translated images used for training the classfication model to evaluate the performance improvement by using this image translation technique for synthetic to real domain adaptation.

2. **RoCoG-v2:** For this video dataset, translation experiments are performed per frame. The "advance" class video is used. The real frames of this video are used for training the diffusion model. This trained model is intergrated with a MultiControlNet model with two conditions: edge and pose. The synthetic frames of this class are translated using this model. The edge maps and pose for each frame are generated and passed to this model along with the prompt *"a person in the style of sks person"*, where ‘sks’ is the learnt placeholder token (identifier).

Image Classification

This section contains the classification experiments on VisDA-2017. A Swin-based transformer (swin-base) pre-trained on ImageNet is used as the backbone. The evaluation metric used in this experiment section is classification accuracy. The baseline experiment is run by training this classifier on all synthetic training images (152,397) and tested on real images (72,372) to provide the lower bound accuracy to compare with the new, translated dataset generated from the tranlation task. The upper bound accuracy is provided by training the classifier with the target domain , validation set images (55,388) and tested on real images.

The translated images mentioned in previous section are then used for training this classifier to evaluate the advantage of using style transfer techniques. This new training set has the number of images as the synthetic training data since all the images are translated per class. The Swin-base transformer is trained on this new training set and tested on real images.

Additional Experiments

This section contains a brief note of the additional experiments performed for the translation task.

1. Synthetic to real translation of the VisDA images using pre-trained SD v1.5 without prompt engineering and integrated with ControlNet (canny version). A simple prompt "a class_name" was utilized. This experiment relies on the prior knowledge of the large generative model.

2. To incorporate the information of target style images, prompts were generated using BLIP [20] captioning model by passing the validation set images. These *"target"* prompts are then mapped to each synthetic image while translation using the same model.

3. A new embedding containing the style of real images is added to this pre-trained diffusion model's text embedding space using Textual Inversion. This new word is added to the *"target"* prompts generated from BLIP for translation.

The results of these experiments can be found in the next chapter.

Chapter 4

Results

Results on VisDA-2017

The translation and classification experiment results on VisDA-2017 are discussed in this section.

Image Translation

A few exampled generated using SD v1.5 model trained on target domain images can be seen in fig 4-1 .

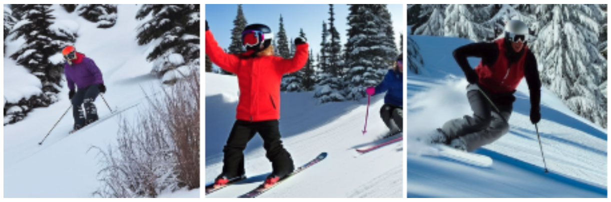


Figure 4-1. Generated examples using prompt: "a person skiing in the style of sks object"

The translation results for each class using our method can be found in 4-2 and fig 4-3. Additionally, comparison of the method with additional experiments can be seen in fig 4-4. Here, the abbreviations are:

- SD+CN+TI+DB: Translation model with Stable Diffusion, ControlNet, Textual Inversion and Dreambooth.

- SD+CN+BLIP+TI: Translation model with Stable Diffusion, ControlNet, BLIP captioning and Textual Inversion.
- SD+CN+BLIP: Translation model with Stable Diffusion, ControlNet, BLIP captioning.
- SD+CN: Translation model with Stable Diffusion, and ControlNet

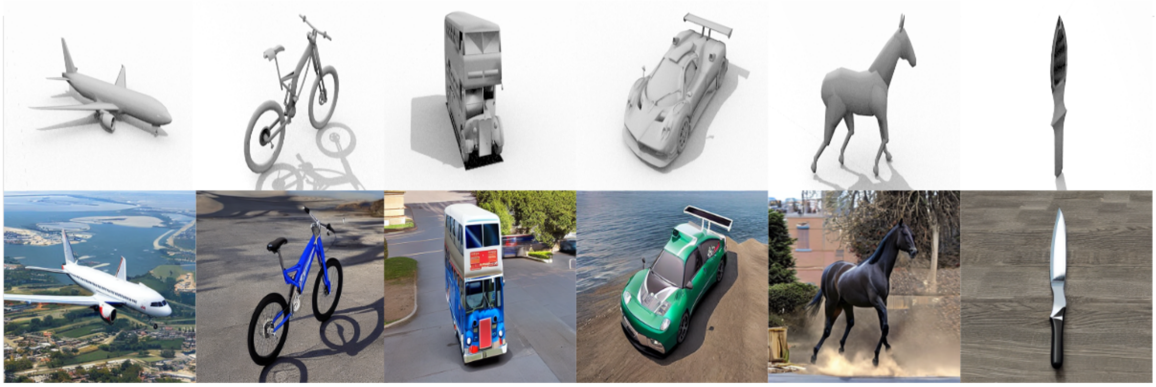


Figure 4-2. Synthetic to Real Image Translation of first 6 classes. (Top) synthetic images and (Bottom) Translated Images

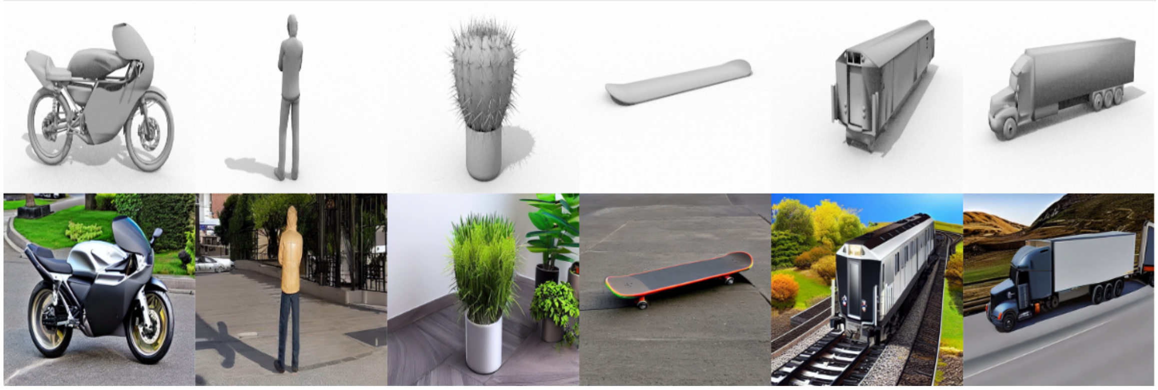


Figure 4-3. Synthetic to Real Image Translation of remaining 6 classes. (Top) synthetic images and (Bottom) Translated Images

Image Classification

The classification accuracy on training the Swin-base classifier with various translated, training sets are given in table 4-I.The abbreviations are explained above.

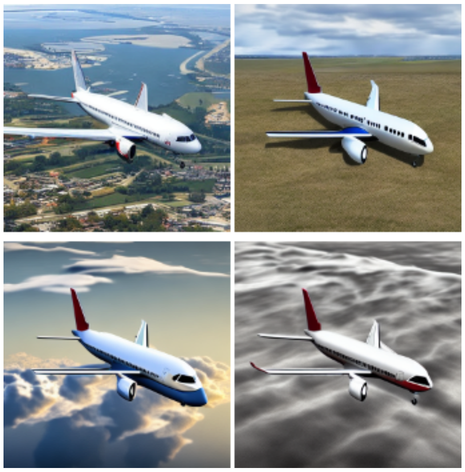


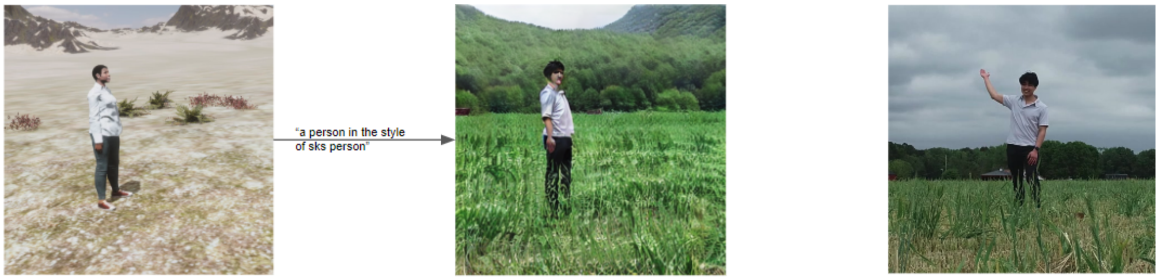
Figure 4-4. Comparison of experiment results. (TL) SD+CN+TI+DB, (TR) SD+CN+BLIP+TI, (BL) SD+CN+BLIP, (BR) SD+CN

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
No DA	75	40	81	**97**	58	16	80	10	77	**56**	**83**	49	62
SD+CN	93	**70**	88	95	77	51	79	**53**	85	28	78	73	75
SD+CN+BLIP	93	52	88	94	74	55	81	49	86	42	80	83	76.76
SD+CN+BLIP+TI	91	59	**89**	96	82	50	87	47	91	49	73	**94**	76.78
SD+CN+TI+DB	**93**	64	86	95	**88**	**58**	**87**	46	**93**	29	72	84	**77.2**

Table 4-I. Classification Accuracy. The best performance is marked as **bold**.

Qualititative Results on RoCoG-v2

Frame to frame translation experiments using one class ("advance") can be seen in fig 4-5.



Synthetic Frame

Translated Frame

Real frame sample used for fine-tuning SD

Figure 4-5. Synthetic to Real Frame Translation of RoCoG-v2

Conclusion and Future Work

An image-to-image translation method for synthetic to real domain adaptation in images, to improve downstream tasks like image classification and action recognition are discussed. Our aim is to translate synthetic images to look more realistic, while preserving the structure of objects. For this task, we first train a diffusion model to generate images with the style of target domain data using Dreambooth and Textual Inversion. This trained model is then integrated with a ControlNet model. Synthetic images are used to generate control images, like edge maps, pose image, etc. These control images along with prompts containing trained tokens are passed to our model to generate translated images with the style of target domain data, which are real world images in our task. Two datasets are used to show results of our model. We observed that the test accuracy on training an image classifier with translated images from our method improved by 15.2% when compared with no adaptation training. This shows that recent advancements in generative models like diffusion models can help in domain adaptation tasks. We also found that this method helps in reducing the randomness in generation by these models, which is a major concern in the field of image recognition. In the near future,

- We aim to extend this method to videos, by adopting temporal consistency in our translation method instead of frame-to-frame translation.
- Investigate the use of language-enhanced prompts and guidance scale for our translation task.

References

1. Gatys, L. A., Ecker, A. S. & Bethge, M. *A Neural Algorithm of Artistic Style* 2015. arXiv: 1508.06576 [cs.CV].
2. Li, Y., Wang, N., Liu, J. & Hou, X. *Demystifying Neural Style Transfer* 2017. arXiv: 1701.01036 [cs.CV].
3. Atapour-Abarghouei, A. & Breckon, T. P. *Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 2800–2810.
4. Shrivastava, A. *et al. Learning from simulated and unsupervised images through adversarial training* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 2107–2116.
5. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics* 2015. arXiv: 1503.03585 [cs.LG].
6. Ho, J., Jain, A. & Abbeel, P. *Denoising Diffusion Probabilistic Models* 2020. arXiv: 2006.11239 [cs.LG].
7. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. *High-Resolution Image Synthesis with Latent Diffusion Models* 2022. arXiv: 2112.10752 [cs.CV].
8. Saharia, C. *et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding* 2022. arXiv: 2205.11487 [cs.CV].
9. Nichol, A. *et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* 2022. arXiv: 2112.10741 [cs.CV].
10. Zhang, L., Rao, A. & Agrawala, M. *Adding Conditional Control to Text-to-Image Diffusion Models* 2023. arXiv: 2302.05543 [cs.CV].
11. Gal, R. *et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion* 2022. arXiv: 2208.01618 [cs.CV].
12. Ruiz, N. *et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation* 2023. arXiv: 2208.12242 [cs.CV].
13. Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V. & Lathuilière, S. *One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 698–708.
14. Wei, P. *et al. Unsupervised Video Domain Adaptation for Action Recognition: A Disentanglement Perspective* 2023. arXiv: 2208.07365 [cs.CV].
15. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M. & Fleet, D. J. *Synthetic Data from Diffusion Models Improves ImageNet Classification* 2023. arXiv: 2304.08466 [cs.CV].

16. Lu, Y., Shen, M., Wang, H., van Rechem, C. & Wei, W. *Machine Learning for Synthetic Data Generation: A Review* 2023. arXiv: 2302.04062 [cs.LG].

17. Schuhmann, C. *et al. LAION-5B: An open large-scale dataset for training next generation image-text models* 2022. arXiv: 2210.08402 [cs.CV].

18. Von Platen, P. *et al. Diffusers: State-of-the-art diffusion models* version 0.12.1. Apr. 2023.

19. Wolf, T. *et al. Transformers: State-of-the-Art Natural Language Processing* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, Online, Oct. 2020), 38–45.

20. Li, J., Li, D., Xiong, C. & Hoi, S. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation* 2022. arXiv: 2201.12086 [cs.CV].