ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CVL Computer Vision Lab

# Robust Object Detection with Efficient Labeled Data Factory

Master's Thesis

Han Sun
Department of Civil, Environmental and Geomatic Engineering

**A**dvisors: Rui Gong
Dr. Yuhua Chen
**S**upervisors: Prof. Dr. Konrad Schindler
Prof. Dr. Luc Van Gool

March 28, 2022

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

---

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

**Title of work** (in block letters):

Robust Object Detection with Efficient Labeled Data Factory

**Authored by** (in block letters):

*For papers written by groups the names of all authors are required.*

| **Name(s):** | **First name(s):** |
|---|---|
| Sun | Han |
| | |
| | |
| | |
| | |

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

| **Place, date** | **Signature(s)** |
|---|---|
| Zurich, 28.Mar.2022 | Han Sun |
| | |
| | |
| | |
| | |

*For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.*

# Abstract

Domain adaptive object detection aims to leverage the knowledge learned from a labeled source domain to improve the performance on an unlabeled target domain. Prior works typically require access to the source domain data for adaptation, as well as the availability of sufficient data in the target domain. However, these assumptions may not hold due to data privacy and rare data collection. In this thesis, we propose and investigate a more practical and challenging domain adaptive object detection problem under both *source-free* and *few-shot* conditions, named as SF-FSDA. To overcome this problem, we develop an efficient labeled data factory based approach. Without accessing the source domain, the data factory renders i) infinite amount of synthesized target-domain-like images, under the guidance of the few-shot image samples and text description from the target domain; ii) corresponding bounding box and category annotations, only demanding minimum human effort, *i.e.*, a few manually labeled examples. On the one hand, the synthesized images mitigate the knowledge insufficiency brought by the few-shot condition. On the other hand, compared to the popular pseudo-label technique, the generated annotations from the data factory not only get rid of the reliance on the source pretrained object detection model, but also alleviate the unavoidably pseudo-label noise due to domain shift and source-free condition. The generated dataset is further utilized to adapt the source pretrained object detection model, realizing robust object detection under SF-FSDA. The experiments under different settings demonstrate that our proposed approach outperforms other state-of-the-art methods on the SF-FSDA problem.

# Acknowledgements

I would like to thank a number of people for their help during this thesis.

- **Prof. Dr. Konrad Schindler** and **Prof. Dr. Luc Van Gool** for offering me the opportunity to work on this topic.
- **Computer Vision Laboratory, ETH Zurich** for the assistance and organization.
- **Rui Gong** for undertaking the supervision and offering continuous support and encouragement during this thesis. He is experienced in this field and is always patient in answering all my questions and suggesting practical solutions. I have learned a lot from our discussions.
- **Dr. Yuhua Chen** for undertaking the supervision and providing guidance through this work. He always provides fresh insights and intriguing exploration directions during our weekly meetings. I strongly benefit from his valuable inputs.
- Last but not least, I would like to thank my parents for their unconditional love and understanding, and my friends for their accompany throughout the whole process.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Object detection, which aims at recognizing and localizing the object instances of certain classes in an image, is a fundamental problem in computer vision. Driven by the rapid development of deep learning and the availability of large-scale datasets, object detection has achieved great advancement over the past decade [61, 48, 60, 7]. However, the performance and generalization ability of the detection system is highly dependent on the availability of manually labeled and diverse datasets, whose labor cost for annotation can be extremely expensive. When applied to the images of which the distribution is different from the training images, the detection models typically exhibit poor generalization, which is common in real applications due to the difference in weather, illumination, object appearance, *etc*.. Thus, recently, domain adaptive object detection problem has been studied [9, 66, 27, 36, 73, 59], which aims to transfer the knowledge learned from the labeled source domain to the unlabeled target domain to train a robust cross-domain object detection model, reducing the effort and cost of human annotation for the target domain.

Generally, existing domain adaptive object detection works reduce the domain shift between the source domain and the target domain, by matching and aligning the source and target domain representations in some space (input space [25, 30, 4] and/or feature space [9, 14]) through the typical techniques of adversarial learning [62, 66], pseudo-label [64, 37, 53], and image translation [28, 30]. They typically assume that, i) the source domain images are accessible when adapting to the target domain, and/or ii) there are abundant images available in the target domain. However, both of these assumptions may not hold in real applications. For example, the data privacy rules and the limited data transmission capacity can break the assumption i), *i.e*., inducing the *source-free* condition, while the rare species image collection and the special medical applications can hinder the assumption ii), *i.e*., causing the *few-shot* condition. The aforementioned domain adaptive object detection techniques can tackle the isolated *source-free* or *few-shot* condition,

but cannot deal with the two conditions at the same time. To be more specific, pseudo-label based techniques are popularly utilized in *source-free* conditions [44, 39], but are not capable of handling *few-shot* conditions, since it relies on enough samples to reduce the pseudo-supervision noise brought by the domain gap. In contrast, adversarial learning [52, 19] and image translation [49, 50] based methods can operate under *few-shot* conditions, but require access to the source domain.

In this work, we study the domain adaptive object detection problem under both *source-free* and *few-shot* conditions, named as SF-FSDA, *i.e.*, the source domain images are not accessible when adapting the object detection model to the target domain, and there are only a few samples available in the target domain (see Fig. 1.1).

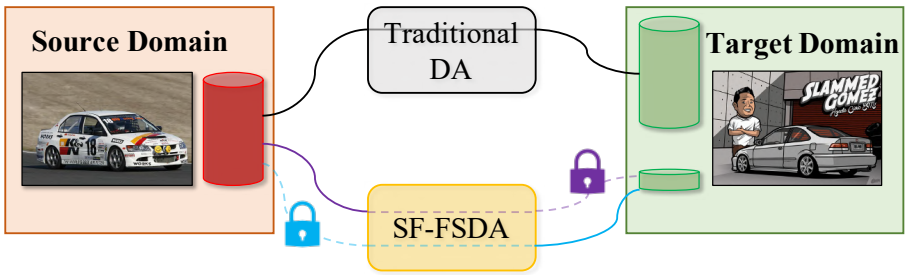


Figure 1.1: Comparison between traditional domain adaptive object detection (DA) problem and our proposed SF-FSDA problem

## 1.1 Focus of this Work

In this thesis, we define the challenging SF-FSDA problem and manage to address this problem via the efficient labeled data factory based approach (see Fig. 1.2), which can automatically generate sufficient target-domain-like images and their corresponding object detection labels, once provided with the text guidance, few-shot image guidance and minimum human annotation. In a nutshell, our work makes the following main contributions:

- We discuss the domain adaptive object detection problem under the *source-free* and *few-shot* conditions, named as SF-FSDA, where there are only a few samples available in the target domain, and only the source pretrained model is accessible for the adaptation to the target domain.
- We develop the efficient labeled data factory based approach, where infinite training data could be synthesized in this restricted condition with minimum human annotation effort.

- We prove the effectiveness of our proposed method for the SF-FSDA problem via experiments on different benchmarks, serving as a strong baseline for further research.

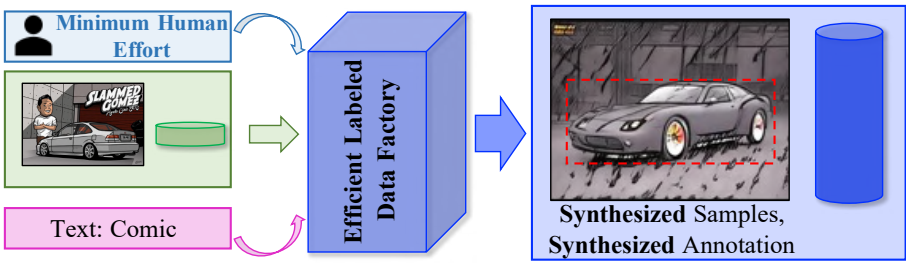


Figure 1.2: Our proposed efficient labeled data factory

## 1.2 Thesis Organization

This Thesis is organized as follows:

- In Chapter 2, we introduce works related to our topic. We mainly cover two parts of correlated research. The first part talks about generative adversarial networks (GANs) with their applications in image style adaptation and dataset synthesis. The second part lies in domain adaptation, focusing on domain adaptive object detection and domain transfer.
- In Chapter 3, we describe our proposed efficient labeled data factory based approach in detail. The problem statement and challenges are defined. We illustrate our general idea for solving the SF-FSDA problem and explain our methodology with the network design.
- In Chapter 4, we present the results of our experiments. We mainly evaluate the performance of our approach in solving the SF-FSDA problem. Ablations to examine the synthesis quality of our data factory are also provided.
- In Chapter 5, we explore other possible solutions and discuss the potential applications and outlooks of our approach.
- In Chapter 6, we draw conclusions of our work.

# Chapter 2

# Related Work

Our efficient labeled data factory based approach relies on synthesized training samples successfully adapted to the target domain and automatically generated annotations for domain adaptation of the downstream object detection task. In this chapter, we introduce a wide range of methods for image style adaptation based on GANs, as well as GAN-based dataset synthesis approaches. Other domain adaptive object detection works are also discussed under the setting of the SF-FSDA problem we proposed.

## 2.1 Generative Models

Recently, GANs [23] have become an active research area and boosted numerous applications, especially image synthesis. It is demonstrated that, given proper training, GANs can synthesize semantically meaningful data from standard data distributions. The current state-of-the-art GAN models [6, 20, 85, 33] are able to generate high-quality realistic images of diverse categories. The rapid development of GANs powers a wide range of applications, including image projection and editing [3, 2, 1, 56, 46], image-to-image translation [88, 55, 29], and domain adaptation [72, 71].

Recent style-based generators [34, 35, 32, 33] produce impressive results and is proven to allow for flexible style control via mapping noise vectors to a higher-dimensional semantic space, which inspires several extensions such as image manipulation [56, 18, 90], image editing [46, 2], and dataset synthesis [86].

## 2.2 Image Style Transfer with GANs

Image style transfer aims to adapt the image in a source domain to a target domain with a different style while keeping the original content. Style adaptation based on conditional generative models has been explored in different directions [5, 88, 56, 18]. In this section, we discuss works focusing on two aspects: image manipulation and image synthesis.

### 2.2.1 Image Manipulation

To transfer images to the desired style, one group of research working on the image-to-image translation task designs a network to directly learn the mapping from the input image to the output one [88, 55]. [88] deploys the cycle consistency loss, which enables training with unpaired images. In [55], they try to further preserve the content when the relationship between the source and target domains is not bijection, and proposes patchwise contrastive learning.

Another group of work explores the GAN latent space, trying to learn disentangled latent representations [51, 16, 79] for image editing or manipulation. [29] is based on the assumption of the latent GAN space to be decomposable. Based on this intuition, they encode the source image into a style space and a shared content space, recombine the code with the content encoding of the source image and random style encoding from the target domain, and then generate the output transferred image with a decoder.

Such GAN inversion idea has been explored more intensively since the recent work of style-based generators [34, 35, 32, 33]. The semantically meaningful latent code space of StyleGAN [34] and StyleGAN2 [35] leverages the possibility of additional style control. Specifically, the general network structure consists of two main components: a mapping network that projects the input noise vector $z$ to the intermediate latent code $w$ in a high-dimensional space $\mathcal{W}$ or $\mathcal{W}+$ for the purpose of disentanglement, and a synthesis network, which progressively generates high-quality images deploying the idea of the previous work [20]. Based on the work of StyleGAN, [1, 68, 63] train an encoder to efficiently embed images into the latent StyleGAN space. The generated latent codes allow for high-quality image reconstruction, which empowers semantically meaningful image manipulation in the latent space. Similar to [29], [63, 90] do style mixing in the latent space by recombining the embedded codes of the content input and the style input, outputting images with higher quality and more detailed control. [81] trains a multi-model encoder, which, instead of a target style image, encodes a target text into a uniform embedded space with the source image, and does style mixing similar to the aforementioned image encoding works. Moreover, with a semantically meaningful space, it is also possible to directly edit and optimize the random latent code to generate desired results. A

number of works manage to correlate semantic features at different scales *e.g.*hair color, pose, to specific editing directions in the latent code space. [24, 67] discover important editing directions by layer-wise decomposition. [56, 46, 2] optimize the latent code with extra supervision from text and/or mask labeling to achieve the desired image manipulation/editing. [91, 79] try to explore an even more disentangled StyleGAN feature space to enable more precise control.

### 2.2.2 Image Synthesis

Another direction explored in GAN style adaptation is image synthesis. Instead of manipulating or editing real images, the GAN network is adapted to generate infinite synthesized images in the target style. The typical approach is fine-tuning a pretrained network by leveraging extra style guidance from the target domain. Various types of training guidance [42, 46, 15, 12, 57] are utilized for controlling the adaptation process, among which the most relevant to our work are text guidance and few-shot image guidance.

The text guidance is typically introduced by learning an image-text multi-model embedding. Recent work by [58] introduces the Contrastive Language-Image Pre-Training (CLIP) model, which learns joint vision-language representations from the large-scale dataset of (image, text) pairs collected from the internet. Based on this powerful pretrained model, [56] incorporates the text guidance for the image manipulation task by mapping the images and the text to the joint embedding space of CLIP. [18] then extends this idea to the image synthesis task by deploying a dual-generator strategy to utilize the image-text pair directional loss for fine-tuning the generator.

The few-shot image guided image synthesis works [65, 47, 54] aim to fully exploit the knowledge from limited training samples, at the same time prevent mode collapse and overfitting. [75] aims to generate images with poses specified by the provided training sketches. They deploy a dual-discriminator strategy and introduce two adversarial loss terms. The cross-domain adversarial loss is to encourage the translated image to look like the few-shot training samples, while the second adversarial loss encourages the generated images to still look like the original ones to deal with the degradation in image quality and diversity. Moreover, they adopt the data augmentation strategy introduced in [32] and only update the mapping network while freezing the other parts of the generator to further prevent overfitting. [54] also introduces loss terms for both few-shot guidance and overfitting prevention. Instead of directly comparing the generated images with the training samples, they compute cross-domain distance consistency across intermediate feature layers. Besides, they deploy the dual-discriminator strategy with discriminators at different levels. [90] explores the extreme case of one-shot image guidance, utilizing pretrained GAN inversion encoder and CLIP embedding. They introduce image guidance by keeping

generating images with the inverted latent code of the guidance image and enforce the generated result to resemble the generated image. In their work, only the synthesis network is updated. [80] discusses the relationship between the pretrained model and fine-tuned model, which indicates the possibility of semantic control for image synthesis task as well.

Different from the aforementioned works, our proposed data factory exploits both the text guidance and the few-shot images guidance together, promoting each other to further improve image synthesis in the target domain.

## 2.3 GANs for Dataset Synthesis

Common works which utilize GANs to synthesize datasets mainly explore the direction of cross-domain adaptation of existing datasets with abundant annotations available. To be more specific, they translate the existing labeled dataset images to obtain annotated dataset in the new domain. [86] is the first work to directly generate training images together with the labels. In this work, they take advantage of the semantically meaningful intermediate feature maps of StyleGAN [34], and upon those features, train simple multi-layer perceptrons to generate semantic segmentation annotations. Another co-concurrent work [43] generates the images and semantic segmentation labels at the same time, and enables image inference by incorporating a GAN inversion encoder.

These two works only focus on the dense prediction task, *e.g.*, semantic segmentation, and do not consider the domain adaptation problem. Instead, the label synthesis branch of our proposed data factory tackles the domain adaptation problem with the few-shot samples and text guidance, and investigates the synthesis of object detection annotations.

## 2.4 Domain Adaptive Object Detection

Domain adaptation aims to transfer knowledge between the label-rich source domain and the unlabeled target domain to train the model that performs well on the target domain. In the past decades, it has been explored in different tasks, *e.g.*, image classification [72, 22, 19], semantic segmentation [70, 74, 69], and object detection [9, 36, 73]. Among the quite vast scope, the most relevant category to our work is domain adaptive object detection, where adversarial learning, image translation, and pseudo-label based methods are typically proposed and studied. Recently, considering more practical scenarios, some works explore the source-free [44] or few-shot [76] domain adaptive object detection problem, respectively. More specifically, [44] tackles the source-free [40, 82, 41] domain adaptive object detection problem with the pseudo-label based technique. And

[76] studies the few-shot [52] domain adaptive object detection problem through adversarial learning based method. However, none of the aforementioned works investigate both the *source-free* and *few-shot* conditions at the same time. In contrast, our SF-FSDA problem touches both *source-free* and *few-shot* conditions simultaneously, which is more challenging and practical. From the method aspect, instead of exploiting pseudo-label or adversarial learning, we synthesize the target domain-like images and the corresponding bounding box and category annotations together with the efficient labeled data factory, without accessing the source domain.

## 2.5 Domain Transfer with Auxiliary Knowledge

In some domain transfer related works, *e.g.*, domain adaptation, domain generalization, and domain randomization, the auxiliary knowledge from the public dataset is utilized as the bridge to connect the source domain and the target domain. For example, since the target domain images are not available for training, [84] randomizes the style of the source domain images utilizing the images from the public dataset ImageNet [13], to improve the generalization ability of the semantic segmentation model trained on the source domain. [78] adopts the auxiliary images from ImageNet to regularize the image classification model training in the adaptation process, to prevent the model from forgetting. However, these works all require access to the auxiliary images, which might not be practical due to data privacy regulations and data transmission capacity. Instead, our efficient labeled data factory takes the publicly available GAN pretrained weights [35] as the auxiliary knowledge, which is more flexible and renders unlimited and unified image and label synthesis.

# Chapter 3

# Methodology

In this chapter, we present our efficient labeled data factory based method to address the SF-FSDA problem, where abundant target-domain-like images with corresponding labels are generated with guidance from the few-shot samples, text description, and few-shot manual annotations. Compared to the existing image translation based approaches [89, 55, 29, 47, 31], our proposed data factory-based method does not require access to the source domain (*source-free condition*), and effectively exploits the knowledge from the few-shot images (*few-shot condition*), together with the text, for image synthesis. Besides, our method goes one step further and synthesizes bounding box annotations for the generated images at the same time, helping improve the downstream domain adaptive object detection task.

For the structure of this chapter, we first set up the problem, and then introduce our proposed solutions under the defined condition. The sections are organized as follows:

1. We discuss the challenges under our specific problem setup, and illustrate our motivation.
2. We present the overall pipeline of our proposed efficient labeled data factory based approach, as shown in Fig. 3.1.
3. We talk about the architecture of the image synthesis branch and the label synthesis branch of our designed network.

## 3.1 Problem Statement

For the problem of domain adaptive object detection, we are given the labeled source domain $\mathcal{S} = \{\mathbf{x}_s^i, \mathbf{y}_s^i\}_{i=1}^{N_s}$ and the unlabeled target domain $\mathcal{T} = \{\mathbf{x}_t^i\}_{i=1}^{N_t}$, where $\mathbf{x}_s^i, \mathbf{y}_s^i$

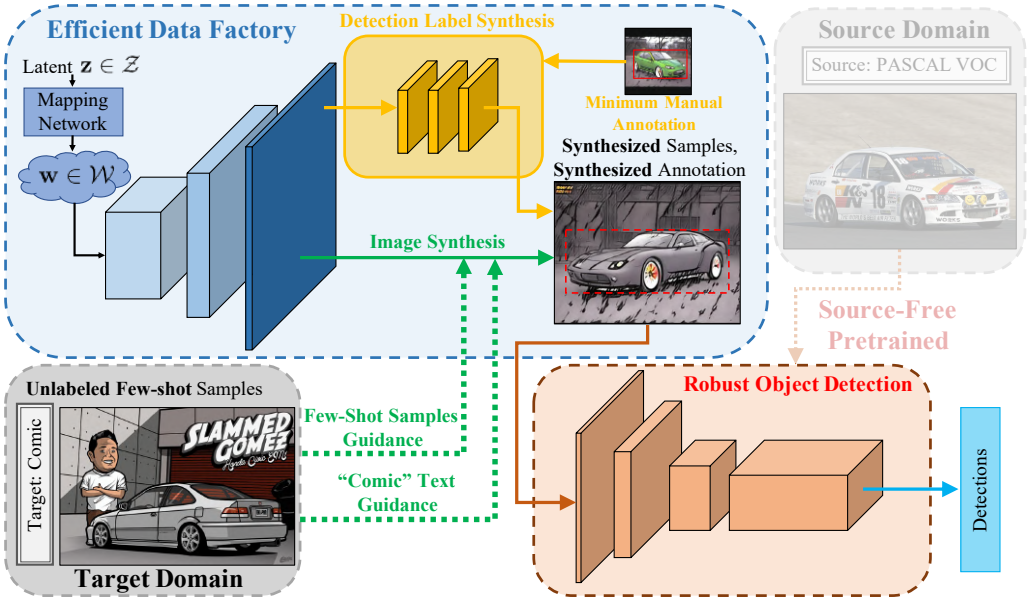


Figure 3.1: Method overview of our efficient labeled data factory based approach

represent the $i$-th image and the corresponding bounding box and category annotations for object detection in the source domain, and $\mathbf{x}_t^i$ denotes the $i$-th unlabeled image in the target domain. $N_s, N_t$ are the number of images in the source and target domain, respectively. Different from traditional domain adaptive object detection problem, we tackle the *source-free* and *few-shot* target conditions, *i.e.*, $N_s \gg N_t$ and $\{\mathbf{x}_s^i, \mathbf{y}_s^i\}_{i=1}^{N_s}$ is not accessible during the adaptation process to $\{\mathbf{x}_t^i\}_{i=1}^{N_t}$, named SF-FSDA problem.

Serving as an example, general pipeline of taking 'Comic' as the target domain is shown in Fig. 3.1. Under this setting, we take the well-labeled PASCAL VOC dataset as the source domain, while there are only few-shot unlabeled Comic samples available in the target domain. The aim is to train the domain adaptive object detection model under the *source-free* and *few-shot* conditions, *i.e.*, only the source pretrained model and few-shot target domain samples are available for the adaptation to the target domain.

### 3.1.1 Technical Challenges

Compared to the traditional domain adaptive object detection problem, our proposed SF-FSDA problem introduces more challenging *source-free* and *few-shot* conditions. Previous techniques for domain adaptive object detection highly rely on adversarial feature learn-

ing [9], image-to-image translation [30], and pseudo-label-based self-training [64]. On the one hand, the challenge brought by the *source-free* condition is that, the previous adversarial feature learning and image-to-image translation-based techniques require access to source data during the adaptation process to align the distribution between the source and target domains, making them not equipped to be engaged in our *source-free* setting. On the other hand, the challenge induced by the *few-shot* condition is that, the pseudo-label based self-training technique always relies on the availability of abundant target domain images to reduce the prediction noise and improve the prediction confidence on the target domain, which are difficult to operate in our *few-shot* setting. Thus, both the *source-free* and *few-shot* conditions hinder the knowledge transfer between the source and target domains for object detection.

### 3.1.2 Motivation

As discussed in the aforementioned technical challenges, the *source-free* and *few-shot* conditions add on difficulty to address the domain adaptive object detection problem. Thus, we aim to firstly adapt on the image level, *i.e.*, synthesize the target-domain-like images. However, different from the previous image translation methods that rely on the access to both the source domain and the target domain, the adaptation is applied on the publicly available trained GAN model, with only few-shot image and text guidance from the target domain. During this process, no access to the source domain training data is required, which provides more flexibility. Moreover, in order to provide reliable guidance for the downstream object detection task, the method for synthesizing the corresponding object detection labels is developed. Inspired by the observation that the trained GAN model encodes the rich knowledge related to the object category and position implicitly in the latent feature space, we introduce the label synthesis branch to produce the object category and bounding box annotation automatically, providing only minimum human effort, *i.e.*, few-shot manual annotation.

## 3.2 Efficient Labeled Data Factory for SF-FSDA Problem

In order to deal with the SF-FSDA problem, we propose the efficient labeled data factory based method, which i) synthesizes abundant target-domain-like images guided by the few-shot samples and the text description from the target domain, without accessing the source domain image; and ii) automatically generates the corresponding object bounding box and category annotations, with the help of minimum human effort, *i.e.*, few-shot manual annotation, as shown in Fig. 3.1.

Since the SF-FSDA problem touches the *source-free* setting, the whole training stage will be divided into, i) source-pretraining stage, ii) image and label synthesis stage and iii) target-adaptation stage. In the i) source-pretraining stage, we get the source pretrained object detection model. The model is trained on the source domain, and this would be the only stage where we leverage the access to the source training data. Then in the ii) image and label synthesis stage, we produce the domain adaptation training data for the original object detection model. The efficient labeled data factory is driven by the few-shot image samples and text guidance from the target domain to synthesize the images with the image synthesis branch, and at the same time automatically synthesizes the corresponding object detection labels by only providing the few-shot manual annotations with the label synthesis branch. In the iii) target-adaptation stage, we adapt the object detection model in the source domain to the target domain. The synthesized images and corresponding labels generated in stage ii) are exploited as training data in the target domain to fine-tune the source pretrained object detection model in stage i).

## 3.3 Image Synthesis

In this section, we talk about the image synthesis branch, where we adapt the GAN model to generate target-domain-like images. Given a publicly pretrained GAN model with the generator $G$, we aim to learn an adapted generator $G_t$ guided by the few-shot image samples $\{\mathbf{x}_t^i\}_{i=1}^{N_t}$ from the target domain $\mathcal{T}$, incorporating text description $T$ simultaneously. The whole network structure is based on StyleGAN2 [35] with additional dual-generator and dual-discriminator design to incorporate both few-shot image and text guidance, and prevent overfitting under the few-shot condition.

The original work in [35] introduces a style-based generator structure, which leverages control of style by introducing a semantically meaningful feature space. As shown in Fig. 3.2, the generator $G$ consists of two main components: the mapping network, which maps the original random noise vector input $z$ to $w$ in higher-dimensional latent code space, and the synthesis network, which follows the previous work [20], progressively generating higher-resolution images during which different levels of styles could be controlled via the intermediate feature maps. In the following content, we introduce how the few-shot image guidance and the text guidance are incorporated, respectively, with an additional freezing strategy for further improvement.

### 3.3.1 Image Synthesis with Few-Shot Image Guidance

To incorporate few-shot image guidance, we follow the work of [54]. The starting point is based on the standard solution of fine-tuning the pretrained GAN model, composed of

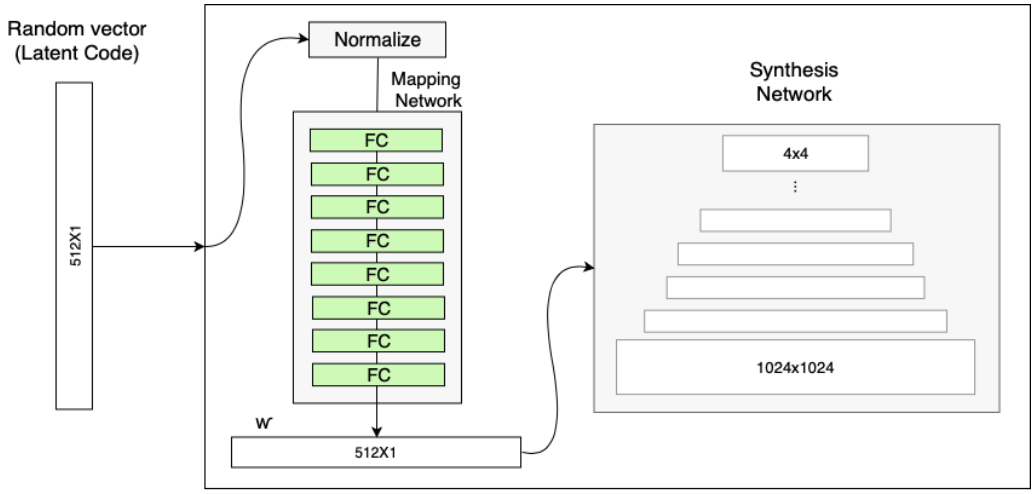


Figure 3.2: General Network Architecture of StyleGAN. Figure from [26]

a generator $G$ and a discriminator $D$, on the few-shot training samples with the common GAN training procedure, which aims at solving the non-saturating objective:

$$\begin{aligned} \mathcal{L}_{\text{adv}}(G, D) &= D(G(z)) - D(x) \\ G_t^* &= \mathbb{E}_{z\sim p_z(z), x\sim \mathcal{D}_t} \arg\min_G \max_D \mathcal{L}_{\text{adv}}(G, D) \end{aligned} \tag{3.1}$$

To address the few-shot condition, the distance consistency regularization, $\mathcal{L}_{dist}$, is utilized to preserve the original content and diversity of the generated images, and the anchor-based relaxed realism is adopted to further prevent overfitting to the few-shot image samples.

**Distance Consistency Regularization** The distance consistency regularization prevents collapse to the few-shot training samples via encouraging the adapted images generated by $G_t$ to still keep the original variety of images generated by $G$. This regularization is achieved by preserving the relative pairwise distance of images generated by specific input vectors. In more detail, we first sample a batch of noise vectors and use their pairwise similarities in feature space to construct probability distributions for each image, where similarity denotes the cosine similarity between generator activations at specific layers. The probability distributions converted from the similarities of the adapted model $G_t$ and the given publicly pretrained model $G$ are encouraged to be uniform by computing KL-divergence across the intermediate layers, as shown in Fig 3.3.

The original implementation of [54] computes the cross-domain distance consistency loss $L_{dist}$ as shown in Eq. 3.2, on four randomly sampled feature layers at each iteration.

To prevent early collapse in style, *e.g.*, color and texture pattern, we relax the regularization during different training phases. Under our relaxed setting, we allow a reasonable extent of object shape adaptation by computing the distance consistency on layers of different levels. In the initial training phases, we compute the distance consistency only of the deep layers (*i.e.*, layers after the 6th) of the generator. After training for certain epochs, we adapt the training strategy and only compare consistency on the shallow layers (*i.e.*, layers before the 10th) for detailed style adaptation and to preserve the content. We deploy this adjustment due to the more significant domain gap to be mapped in our case. Experiment in the original paper shows that, in most cases, they fail to adapt the source domain to the target domain when their contents are unrelated (see Fig 3.4). We show successful adaptation of the original model to the unrelated target domain and observe minor collapse in style *i.e.*, the same color or pattern before successful adaptation under our setting.

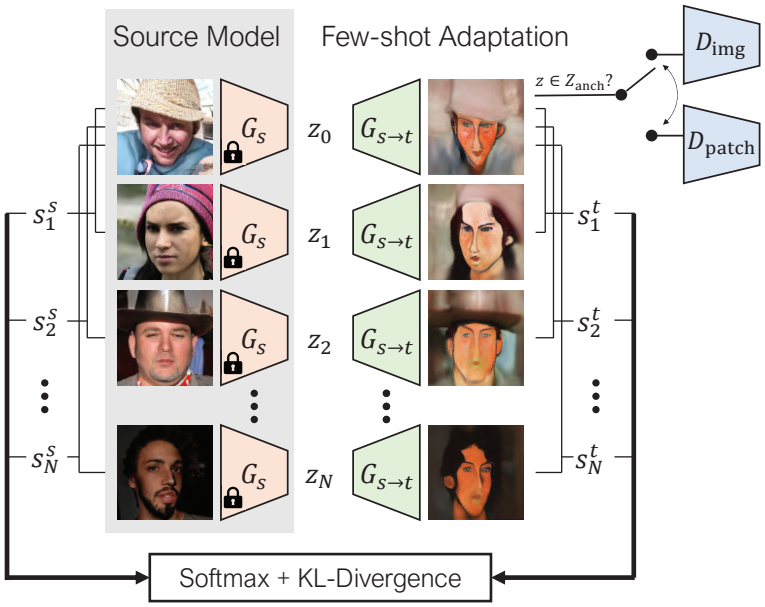


Figure 3.3: Cross-domain consistency loss and relaxed realism approaches. Figure from [54, Figure 3]

**Relaxed Realism** In order to further prevent the synthesized images from overfitting to the few-shot image samples of the target domain, a dual-discriminator strategy is deployed for introducing the relaxed realism [54]. The general idea is to only discriminate the fake and real images on the image level when the input vector is sampled from a limited region

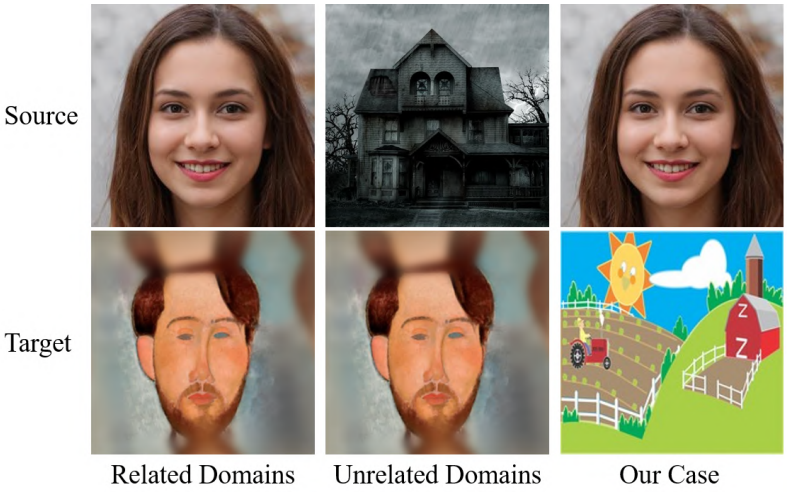


Figure 3.4: Explanation of related domains and unrelated domains

while allowing other generations to resemble the sample images only on the patch level. In detail, firstly, an anchor region is defined as a subset of the entire input latent space $\mathcal{Z}$. When sampled from these regions, we use a full image discriminator $D_{img}$. Outside of them, we enforce adversarial loss using a patch-level discriminator $D_{patch}$ to avoid overfitting to the few-shot samples.

The image-level discriminator $D_{img}$ follows the original design as in [35], while the second discriminator $D_{patch}$ is defined as the subset of the original discriminator $D_{img}$. In more detail, 4 extra convolutional layers are defined for the $D_{patch}$, which reduce the channel of the input to 1. When the input vector is sampled outside the anchor region, the output of the patch-level discriminator $D_{patch}$ is acquired by randomly taking one intermediate feature map of the original discriminator $D_{img}$ and feeding it forward to a random channel multiplier of the extra layers. The two discriminators are deployed at a designed frequency for the purpose of preserving image diversity while still leveraging whole-image guidance. The process is controlled with the sampling frequency hyper-parameter $\lambda_f$, which indicates the frequency of sampling from the anchor region and computing the image-level loss instead of the patch-level loss.

With the aforementioned distance consistency regularization and relaxed realism strategies, the objective of image synthesis with few-shot image guidance is defined as:

$$G_t^* = \arg\min_{G_t} \max_{D_{\text{img}}, D_{\text{patch}}} \mathcal{L}_{\text{adv}}\left(G_t, D_{\text{img}}, D_{\text{patch}}\right) + \lambda_1 \mathcal{L}_{\text{dist}}\left(G_t, G\right), \tag{3.2}$$

where $\mathcal{L}_{\text{adv}}$ represents the adversarial loss, and $\lambda_1$ is the hyper-parameter to balance the adversarial loss and the distance consistency regularization loss.

### 3.3.2 Image Synthesis with Text guidance

Besides the few-shot image samples from the target domain, the text description about the target domain is available with no effort required, *e.g.*"cartoon" and "watercolor." Incorporating text knowledge in our case not only introduces a more flexible way to define adaptation style, but also helps with escaping early collapse to few-shot samples by leveraging extra guidance.

To fully exploit and transfer the knowledge from the target domain to imitate its distribution, text guidance from the target domain can be leveraged to guide the image synthesis of the data factory with the help of CLIP models [58]. The main idea is to train the GAN model to make the generated images shift along the direction of the textually-described path in the CLIP embedding space [18]. Original and target texts are both self-defined to provide the desired shifting guidance. In order to obtain the image shifting direction during the training process, a dual-generator strategy is also deployed. We fix the pretrained generator $G$ to keep generating original images for comparison while optimizing the target generator $G_t$. Then the changing directions of the text guidance and images can be expressed by,

$$\begin{aligned} \Delta T &= E_{text}\left(T_{\text{target}}\right) - E_{text}\left(T\right) \\ \Delta I &= E_{img}\left(G_t(\mathbf{z})\right) - E_{img}\left(G(\mathbf{z})\right), \end{aligned} \tag{3.3}$$

where $E_{text}$ and $E_{img}$ denote CLIP text and image encoders, respectively. $T$ and $T_{\text{target}}$ represent the text description of the pretrained GAN model and the target domain, *e.g.*, "photo" and "comic." $\mathbf{z}$ is the input noise variable, *i.e.*, $\mathbf{z} \in \mathcal{Z}$. The directional loss introduced by text guidance can thus be described as,

$$\mathcal{L}_{\text{direction}}(G, G_t, T, T_{\text{target}}) = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|}. \tag{3.4}$$

Combined with the few-shot image guidance training objective in Eq.(3.2), our final training objective with both the few-shot image guidance and the text guidance can be derived as,

$$\begin{aligned} G_t^* = \arg\min_{G_t} \max_{D_{\text{img}}, D_{\text{patch}}} & \mathcal{L}_{\text{adv}}\left(G_t, D_{\text{img}}, D_{\text{patch}}\right) \\ & + \lambda_1 \mathcal{L}_{\text{dist}}\left(G_t, G\right) \\ & + \lambda_2 \mathcal{L}_{\text{direction}}\left(G_t, G, T, T_{\text{target}}\right), \end{aligned} \tag{3.5}$$

where $\lambda_2$ is the hyper-parameter to balance the text guidance and other terms.

### 3.3.3 Image Synthesis Training Strategy.

In order to further prevent the model from overfitting the few-shot image samples, we adopt the freezing strategy during the training. The shallow layers of the generator are frozen to preserve the original contents of the pretrained model, while shallow layers of the discriminator are frozen accordingly to further ensure a stable training process.

To be more specific, we adapt the generator $G_t$ only on the specifically chosen intermediate convolutional feature layers while freezing the rest part of the network. For all the experiments, we only update the weights of intermediate layers from the third to the last one. Accordingly, we freeze the image-level and the patch-level discriminators, $D_{img}$ and $D_{patch}$, except the final layer. The training strategy is simple yet proven effective in preventing overfitting under few-shot conditions.

## 3.4 Label Synthesis

We can now get unlimited target-like samples with a successfully adapted image synthesis branch. Previous research has proved that StyleGAN2 [35] learns a well-disentangled semantic latent space, where each channel controls meaningful properties at different scales. Intuitively, feature maps generated by those channels should be semantically informative enough to act as extracted features for downstream tasks, e.g., segmentation and detection.

Based on this assumption, we develop our object detection branch by utilizing the adapted StyleGAN2 [35] generator $G_t$ acquired in the image synthesis step as the backbone network for feature extraction, and then incorporating prediction heads on this basis. The general label synthesis branch archirtecture is shown in Fig. 3.5. We get our training data with the following procedure: We sample a set of latent codes $\{z_i\}_{i=1}^{N_a}$ and generate their corresponding images $\{G_t(z_i)\}_{i=1}^{N_a}$ with the image synthesis branch. Here $N_a$ denotes the number of manual annotations required to train the object detection branch. Then we manually annotate these samples as our training data. During the training process, we deploy the generator $G_t$ as our backbone network and take the intermediate convolutional feature maps generated by the latent codes as our encoded features for the matching images. For the detailed implementation, we take the intermediate feature maps with the resolutions (4, 8, 16, 32, 64) considering the memory consumption. Then we upsample those feature maps with bilinear interpolation to the resolution of 128, and concatenate them together to feed forward to the prediction heads.

A prediction network is built on these extracted features for the object detection task. Inspired by [87], we use keypoint representations where each object is represented by its center point and the size of its bounding box. To detect objects presented in a synthesized image $\bar{\mathbf{x}}_t \in R^{W \times H \times 3}$, our goal is to predict a downsampled keypoint heatmap

$\hat{\mathbf{y}} \in [0,1]^{\frac{W}{r} \times \frac{H}{r} \times C}$. $C$ denotes the number of classes for the prediction task, $r$ represents the downsampling stride, and $W, H$ are the width and height of the image. A prediction $\hat{\mathbf{y}}_{x,y,c} = 1$ represents a detected keypoint of class $c$, while $\hat{\mathbf{y}}_{x,y,c} = 0$ means background. For loss propagation, ground truth heatmap $\mathbf{y}$ is generated by converting each ground truth keypoint $p \in \mathcal{R}^2$ to its low-resolution equivalent $\tilde{p} = \left\lfloor \frac{p}{r} \right\rfloor$ and splatting those points using a Gaussian Kernel. The training loss is defined as a variant of focal loss [45],

$$\mathcal{L}_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - \hat{\mathbf{y}}_{xyc})^{\alpha} \log(\hat{\mathbf{y}}_{xyc}) & \text{if } \mathbf{y}_{xyc} = 1 \\ (1 - \mathbf{y}_{xyc})^{\beta} (\hat{\mathbf{y}}_{xyc})^{\alpha} \log(1 - \hat{\mathbf{y}}_{xyc}) & \text{otherwise,} \end{cases} \tag{3.6}$$

where $\alpha$ and $\beta$ are hyper-parameters of the focal loss, while $N$ is the number of keypoints in image $\bar{\mathbf{x}}_t$ for normalization.

A local offset $\hat{\mathbf{o}} \in \mathcal{R}^{\frac{W}{r} \times \frac{H}{r} \times 2}$ is predicted and shared among all classes to recover the precise center point locations in compensation for the error caused by downsampling. The sizes of bounding boxes $\hat{\mathbf{s}} \in \mathcal{R}^{\frac{W}{r} \times \frac{H}{r} \times 2}$ of each class $c$ are regressed around the predicted center points, using a single shared prediction as well. Offset loss is computed only at locations of predicted keypoints $\tilde{p}$, while size loss is computed for each detected object $k$ with its predicted size $\hat{\mathbf{s}}_{p_k}$ around the center point $p_k$ and the ground truth bounding box size $\mathbf{s}_k$. Both keypoint offset and size predictions are trained with L1 loss,

$$\begin{aligned} L_{off} &= \frac{1}{N} \sum_{p} \left| \hat{\mathbf{o}}_{\tilde{p}} - \left( \frac{p}{r} - \tilde{p} \right) \right| \\ L_{size} &= \frac{1}{N} \sum_{k=1}^{N} \left| \hat{\mathbf{s}}_{p_k} - \mathbf{s}_k \right|. \end{aligned} \tag{3.7}$$

Three prediction heads are built for predicting $\hat{\mathbf{y}}$, $\hat{\mathbf{o}}$, and $\hat{\mathbf{s}}$, respectively. Each prediction head is composed of, 3×3 convolutional layer, ReLU, and 1×1 convolutional layer. The prediction heads are trained with a weighted sum of loss terms for these tasks,

$$\mathcal{L}_{det} = \mathcal{L}_k + \lambda_{off} \mathcal{L}_{off} + \lambda_{size} \mathcal{L}_{size}, \tag{3.8}$$

where $\lambda_{off}$ and $\lambda_{size}$ represent the hyper-parameters to balance the offset, size and keypoint prediction training loss.

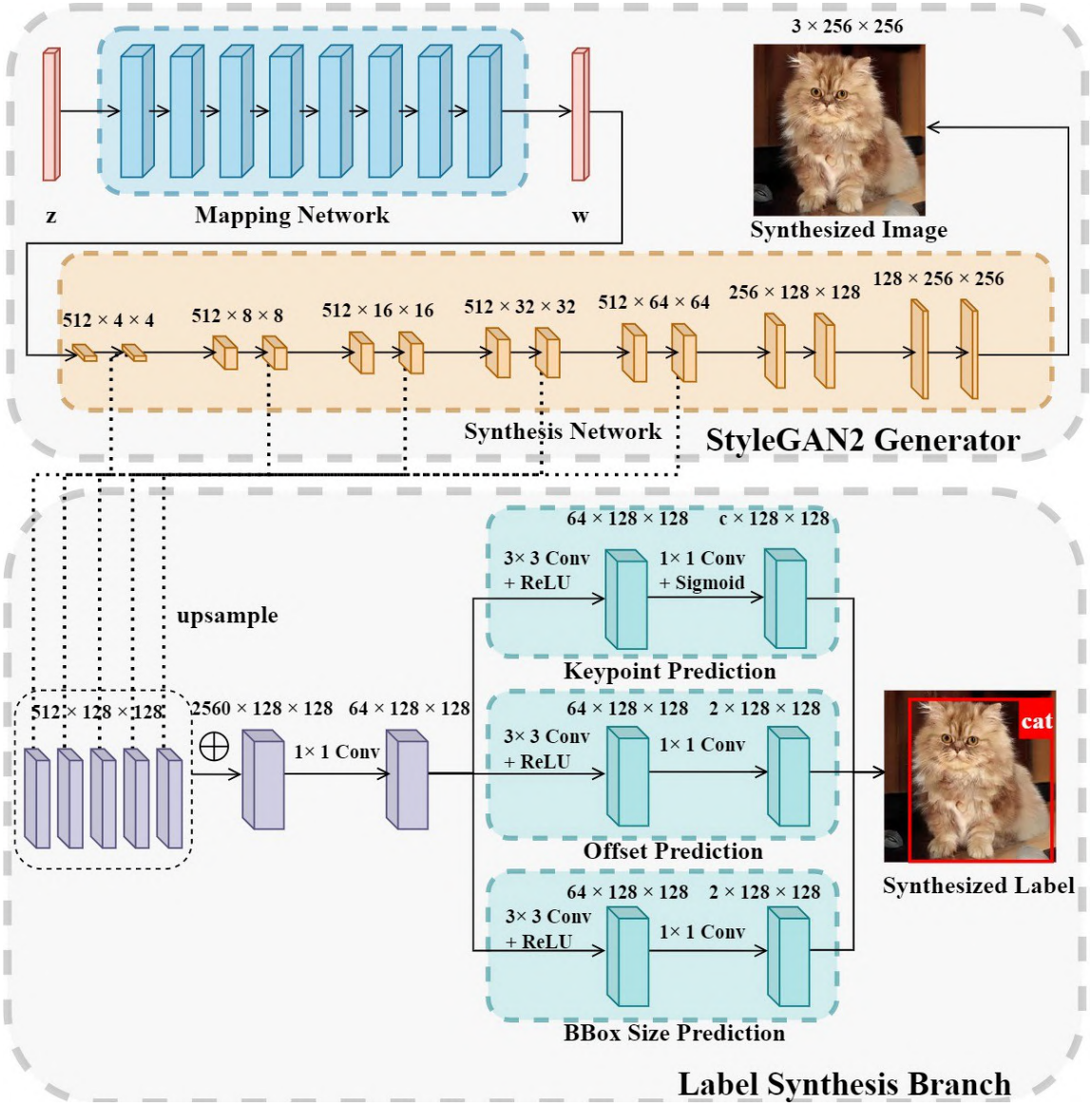


Figure 3.5: Network structure of the label synthesis branch of our efficient labeled data factory, taking image generation with resolution of 256 as an example

# Chapter 4

# Experiments and Results

In this chapter, we first introduce our experimental setup and data sources. Then in the results, we present experiments validating the benefit of our proposed efficient labeled data factory based method in the downstream object detection task, and compare it with other baseline methods. Next, we conduct the ablation study to discuss the supervision introduced by both the few-shot image guidance and the text guidance, and talk about the extra regularization achieved by our freezing strategy. More visual results are provided to prove the effectiveness of our proposed method for image style adaptation as well as detection results on a dataset with multiple classes to show the potential of the label synthesis branch. Finally, we compare our proposed efficient labeled data factory based method with the pseudo-label based one.

## 4.1 Experimental Setup

In order to prove the effectiveness of our efficient labeled data factory for robust object detection, the in-domain and cross-domain experiments are conducted.

### 4.1.1 In-Domain Experiments

In order to verify the validity of our proposed efficient labeled data factory for automatically producing the images and the corresponding object category and bounding box labels, we conduct the in-domain experiments, where our efficient labeled data factory is used to generate the images and object bounding box and category annotations, *without domain adaptation*. Then, the generated images and labels are exploited to train the object detection model, to recognize the object instances in the same or similar domain, *i.e.,* in-domain object detection. More specifically, in our experiment, the object detection model

is trained on the natural images and annotations (see Fig. 4.2(c)) synthesized by our data factory, and tested on the PASCAL VOC dataset.

### 4.1.2 SF-FSDA Cross-Domain Experiments

For the purpose of proving the helpfulness of our proposed efficient labeled data factory for domain adaptation, the SF-FSDA cross-domain experiments are explored, where the data factory is trained with the guidance of the *text* and/or the *few-shot samples* from the target domain. Furthermore, the synthesized images and labels are utilized to fine-tune the source domain pretrained object detection model, to adapt the model to the target domain, *i.e.,* SF-FSDA cross-domain object detection. In our experiments, we aim at realizing SF-FSDA, under PASCAL VOC (source) $\rightarrow$ Clipart and Comic (target), respectively.

## 4.2 Dataset

### 4.2.1 PASCAL VOC

PASCAL VOC 2007 & 2012 datasets [17] contain natural objects with manual annotations. Each image in PASCAL VOC dataset includes the object class, pixel-level semantic label, and object bounding box annotations, serving as an important benchmark for the image classification, semantic segmentation and object detection tasks. Our experiment is related to the object detection task on PASCAL VOC dataset. In the in-domain experiments, the test set with cat and car objects is utilized to evaluate the performance of the object detection model. In the SF-FSDA cross-domain experiments, the training set including labeled car and cat images is taken as the source domain for training.

PASCAL VOC dataset customizes the license, especially the images collected from the Flickr website, *i.e.*, PASCAL VOC dataset grants the limited, non-transferable, non-sublicensable, revocable license to access and use the data.

### 4.2.2 Clipart1k

Clipart1k dataset [30] includes the clipart images collected from the CMPlaces dataset [8] and two image search engines [30]. It covers clipart images, exhibiting a large domain shift compared to PASCAL VOC dataset. In the SF-FSDA cross-domain experiments, 12 unlabeled images are exploited as the few-shot target samples for training, and the test set containing cat and car objects is taken for the model performance evaluation on SF-FSDA.

### 4.2.3 Comic2k

Comic2k dataset [30] covers the comic images collected from BAM! [77]. It consists of comic images, indicating a clear domain gap compared to PASCAL VOC dataset. In the SF-FSDA cross-domain experiments, 5 unlabeled images are regarded as the few-shot target domain for training, and the test set containing cat and car objects is adopted for the model performance evaluation on SF-FSDA.

### 4.2.4 Watercolor2k

Watercolor2k dataset contains the watercolor images collected from BAM! [77]. In the additional SF-FSDA cross-domain experiments, we utilize 5 unlabeled images from this dataset as image style guidance for training.

## 4.3 Implementation Details

### 4.3.1 Image Synthesis

The data factory is based on the StyleGAN2 structure and initialized with the publicly available cat and car image synthesis pretrained weights in [35].

In the SF-FSDA cross-domain experiment, PASCAL→Clipart, we take 12 images from Clipart1k as few-shot image training samples. The source text is "Photo", and the target text is "Clipart." We set the weight of the few-shot image guidance $\lambda_1$ to 1.0 and the weight of the text guidance $\lambda_2$ to 1.0 in Eq. (3.5). In the PASCAL→Comic setting, we take 5 images from Comic2k for image guidance, and the text guidance is defined as "Photo"→"Comic." $\lambda_1, \lambda_2$ are set as 1.0, 5.0, respectively.

For the freezing training strategy described in Sec. 3.3.3, the weights of all tRGB layers [35] are fixed along with the convolutional feature layers with the lowest resolution. All layers except the final layer of the discriminators are frozen.

For the setting of the distance consistency regularization, as described in Sec. 3.3.1, $L_{dist}$ is first computed by sampling from intermediate feature layers after 6. After training for 600 iterations, we adapt the strategy and sample from shallow layers from the first to the 10th.

Another parameter is the sampling frequency from the anchor region, $\lambda_f$, which decides how often we compute the discriminator loss on the whole image level as introduced in the relaxed realism in Sec. 3.3.1. Due to the larger domain gap introduced by style images from unrelated domains in our settings, we set this hyper-parameter to 2, alternatively

computing the loss on the image level and the patch level. The rest training details follow the StyleGAN2 [35] with the augmentation strategy introduced in [32]. The training iteration for image synthesis is set as 1000.

### 4.3.2 Label Synthesis

As the minimum human effort, we manually label 10 synthesized images. The intermediate features in StyleGAN2 used for the label synthesis branch are obtained by taking the second convolutional feature layer with the five lowest resolutions (4, 8, 16, 32, 64), upsampling them, and then concatenating them together, as shown in Fig. 3.5. $\alpha, \beta$ in Eq. (3.6) are set as 2, 4. We set hyper-parameters $\lambda_{off}, \lambda_{size}$ of the loss in Eq. (3.8) as 1.0, 0.5. We adopt the SGD optimizer for training, with the learning rate as 0.0001 and the weight decay as 0.0001. Keypoints are predicted on a heatmap with the resolution of 128. The training iteration for label synthesis is set as 1000.

### 4.3.3 Source Pretraining and Target Adaptation

The object detection model in the source pretraining and target adaptation stage is based on the Single Shot MultiBox Detector (SSD) [48] model. For the in-domain experiments, we deploy the ImageNet pretrained backbone and generate 200 training samples in realistic style together with annotation with our data factory, incorporating the data augmentation strategy introduced in [21]. For the SF-FSDA cross-domain experiments, we synthesize 250 samples with annotations in the desired target style for adaptation, utilizing the same augmentation strategy.

### 4.3.4 Baseline Setup

In Table 4.1, Table 4.2, and Table 4.3, the "Few-Shot FT" represents that the object detection model is fine-tuned on a few images with manual annotations, where the images are generated by the image synthesis branch of our data factory. In Table 4.2 and Table 4.3, the "CycleGAN", "MUNIT" and "CUT" conduct the corresponding image translation methods between the synthesized images from the pretrained StyleGAN2 model and the few-shot target domain images to generate the target-domain-like images, and adopt the same annotations generated by our data factory. Oracle performance in Table 4.1 is reached by training the object detection model on the training set of PASCAL VOC. Oracle performance in Table 4.2-4.3 is obtained by [30] for the traditional domain adaptive object detection, which is neither few-shot nor source-free.

## 4.4 Results

In this section, we present and discuss our experimental results under various settings. We first prove that our proposed model effectively synthesizes the image samples and the corresponding object bounding box and category labels by implementing the in-domain experiment. Then we show that our proposed data factory mitigates the source and target domain gap through the guidance of text and few-shot target domain image examples via the quantitative and qualitative SF-FSDA cross-domain experimental results. Ablations and other extensions are also conducted.

### 4.4.1 In-Domain Experiments

As shown in Table 4.1 and Fig. 4.2(c), our synthesized images and corresponding bounding box labels can be used to train the model for the object detection on the same/similar domain, improving the few-shot object detection performance from 50.86%, 41.57% to 64.37%, 52.73% on the "Cat" and "Car" objects detection, respectively. It opens up a new avenue for the few-shot object detection task, by manually labeling object bounding boxes in a few images, synthesizing enough image samples and bounding box labels automatically with our proposed efficient labeled data factory, and then training the object detection model with the synthesized images and labels.

| | Few-Shot FT | Ours | Oracle |
|---|---|---|---|
| Cat | 50.86 | **64.37** | 86.48 |
| Car | 41.57 | **52.73** | 72.18 |

Table 4.1: In-domain experiments on PASCAL VOC. The results are reported on average precision (AP)

In order to further figure out the effect of the number of synthesized images and annotations from the efficient labeled data factory, the object detection performance with different numbers of synthesized images and samples are shown in Fig. 4.1. It is shown that the object detection performance improves as more images and annotations are synthesized.

### 4.4.2 SF-FSDA Cross-Domain Experiments

In Table 4.2, and Table 4.3, the quantitative results are shown on the benchmark, PASCAL VOC→ Clipart, Comic, respectively. Compared with the pure source baseline, all of

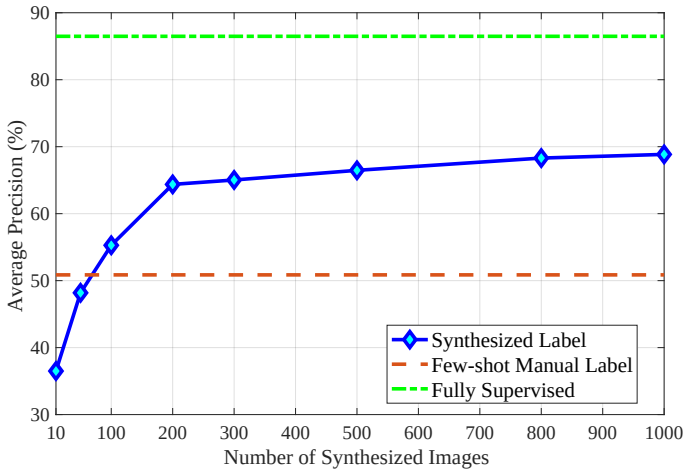


Figure 4.1: Object detection performance with different numbers of synthesized images and annotations, under the setting of Table 4.1

the image style adaptation based methods bring performance improvement, verifying the benefits of the style adaptation based methods for narrowing the domain gap. Among the image style adaptation based methods, it is shown that our proposed data factory based method surpasses other image translation based methods, CycleGAN [89], MUNIT [29], and CUT [55]. It proves the advantage of our method for synthesizing the target-domain-like images, with the guidance of both the few-shot samples and the text knowledge.

| | Source | Few-Shot FT | CycleGAN | MUNIT | CUT | Ours | Oracle |
|---|---|---|---|---|---|---|---|
| Cat | 17.25 | 30.94 | 27.01 | 21.46 | 24.57 | **32.50** | 35.07 |
| Car | 43.04 | 52.97 | 55.11 | 54.62 | 54.72 | **55.67** | 57.38 |

Table 4.2: SF-FSDA cross-domain experiments, PASCAL VOC → Clipart

| | Source | Few-Shot FT | CycleGAN | MUNIT | CUT | Ours | Oracle |
|---|---|---|---|---|---|---|---|
| Cat | 16.36 | 33.01 | 23.51 | 37.28 | 36.81 | **37.74** | 39.99 |
| Car | 39.02 | 51.05 | 42.20 | 41.31 | 46.68 | **54.68** | 52.76 |

Table 4.3: SF-FSDA cross-domain experiments, PASCAL VOC → Comic

Fig. 4.2 presents the qualitative results of our method compared to the baseline methods. (a)-(b) are the exemplar images from the Clipart1k and Comic2k datasets. (c) are

the synthesized images from the publicly available pretrained GAN weights, used in Table 4.1 without conducting domain adaptation. It is notable that our data factory does not have the requirement of on which style images the GAN model is pretrained, and we just adopt the publicly available pretrained weights provided in [35]. (d) are the synthesized image and annotations from our proposed data factory in Table 4.2-4.3. (e)-(g) are the results generated by other image translation methods in Table 4.2-4.3. Although provided with image guidance with unrelated content, our approach learns the general style of the exemplar images appropriately without overfitting and outperforms the other methods obviously. Fig. 4.3 shows the FID scores of the style-adapted images synthesized by our data factory or translated by the other baseline methods, which proves the better image quality of our approach quantitatively.

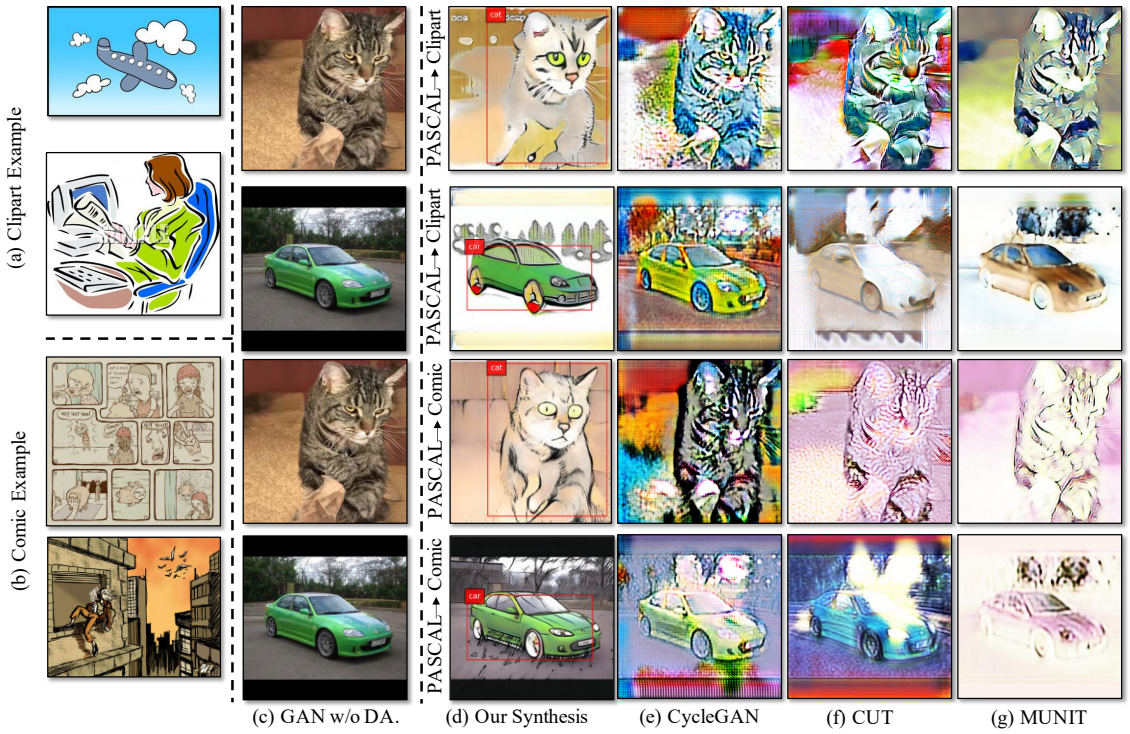


Figure 4.2: SF-FSDA: PASCAL VOC→Clipart & Comic qualitative results

Moreover, as shown in Table 4.2 and Table 4.3, compared with the few-shot manual annotations, the automatically synthesized annotations can further improve the performance. On the PASCAL VOC→ Clipart benchmark, the AP is improved from 30.94%, 52.97% to 32.50%, 55.67%. On the PASCAL VOC→ Comic benchmark, the performance is

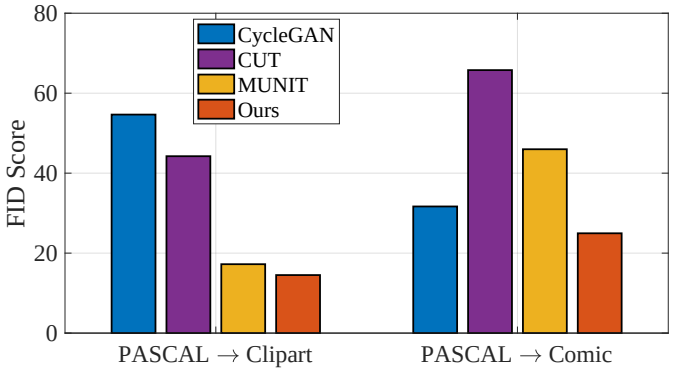


Figure 4.3: Synthesized images quality comparison between our proposed data factory and other image translation based methods, measured with FID score (↓)

improved from 33.01%, 51.05% to 37.73%, 54.68%. It verifies the effectiveness of the automatically generated images and annotations for the SF-FSDA problem.

Besides the two benchmark settings, we conduct the additional SF-FSDA experiments under the PASCAL VOC→Watercolor setting. Compared to the few-shot fine-tuning baseline as done in Table 4.2 and Table 4.3, our proposed efficient labeled data factory based method further improves the object detection performance from 49.03%, 64.47% to 52.06%, 65.56% for the cat and car categories, respectively. In Fig. 4.4, we show the image and label synthesis results from our data factory, under the PASCAL VOC→Watercolor setting.

In Fig. 4.5, we show the qualitative object detection results on the target domain, *i.e.*, Clipart1k and Comic2k. "Before Adaptation" represents the object detection results when applying the source-pretrained object detection model to the target domain. "After Adaptation" shows the object detection results after fine-tuning the source-pretrained model on the synthesized images and labels from our proposed data factory. The image without any detected bounding box indicates that the model cannot detect the objects in the image. Better detection results are generated after cross-domain adaptation based on our method.

### 4.4.3 Ablation Study

**Text and Few-shot Image Guidance** In our proposed efficient labeled data factory for SF-FSDA, the style of the generated samples is guided by the few-shot image samples and/or the text guidance. In order to explore the effect of the two different types of guidance, we compare the performance of different ablations of the full model. From the quantitative comparison in Table 4.4, it is shown that both the few-shot samples and text guidance con-

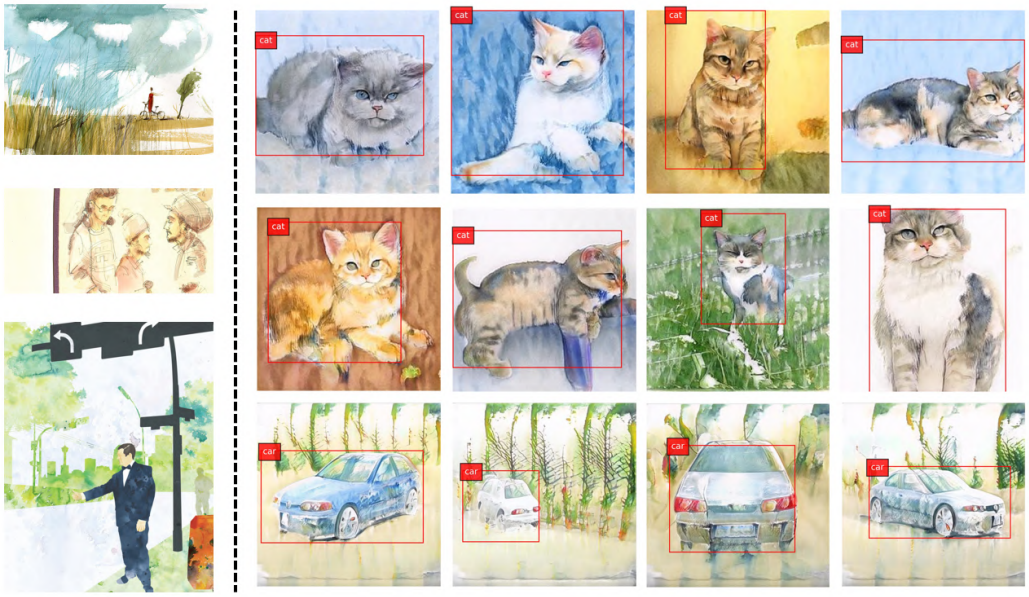


Figure 4.4: SF-FSDA: PASCAL VOC→Watercolor qualitative results

tribute to the final image synthesis results. From the qualitative results shown in Fig. 4.6, taking the "clipart" style as the example, the text guidance provides the general knowledge on what the "clipart" images look like, while the few-shot image guidance indicates how the "clipart" images are on the target domain. Moreover, the text knowledge from the target domain prevents overfitting to the few-shot samples. On the other hand, it is proven that our model is flexible, still reaching effective synthesis results even when one of the text and few-shot samples guidance is not available.

| Source | Only Few-Shot | Only Text | Few-shot+Text |
|---|---|---|---|
| 17.25 | 28.24 | 18.60 | **32.50** |

Table 4.4: Ablation study for the text and few-shot image guidance from the target domain, measured with AP performance on Clipart

**Freezing Strategy** The ablation study on the freezing strategy during training is conducted, which is shown in Fig. 4.7 and Table 4.5. It is shown that the freezing strategy for the image synthesis training can help prevent overfitting to the few-shot samples in the target domain and preserve the diversity of the image synthesis results.


Figure 4.5: Qualitative object detection results on the target domain, Clipart and Comic

In order to further prove the validity of the freezing strategy under the extreme case, we here provide the qualitative comparison in Fig. 4.8 under the one-shot target domain condition, *i.e.*, there is only one image available on the target domain. From Fig. 4.8, it is shown that the freezing strategy is especially important for improving the image generation diversity and preventing overfitting to the one-shot image samples under the challenging one-shot condition.

| w/o freezing | w freezing |
|---|---|
| 0.64 | **0.68** |

Table 4.5: Ablation study for freezing strategy during image synthesis training, measured with the LPIPS distance [54](↑)


Figure 4.6: Qualitative results comparison, with/without text/few-shot image guidance for image synthesis training

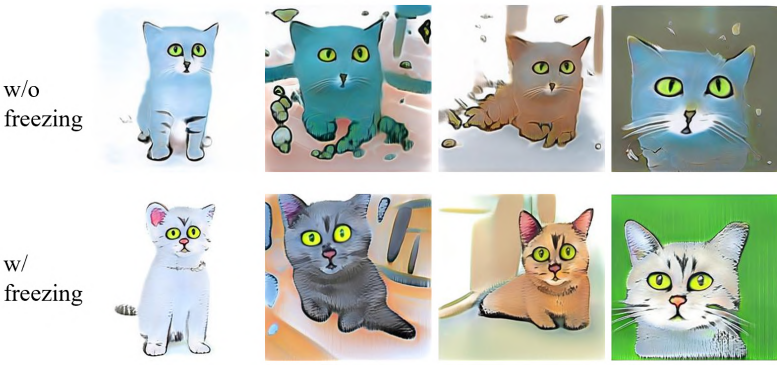


Figure 4.7: Qualitative results comparison, with/without freezing strategy for image synthesis training


Figure 4.8: Qualitative results comparison, with/without freezing strategy for image synthesis training, under the one-shot target domain condition

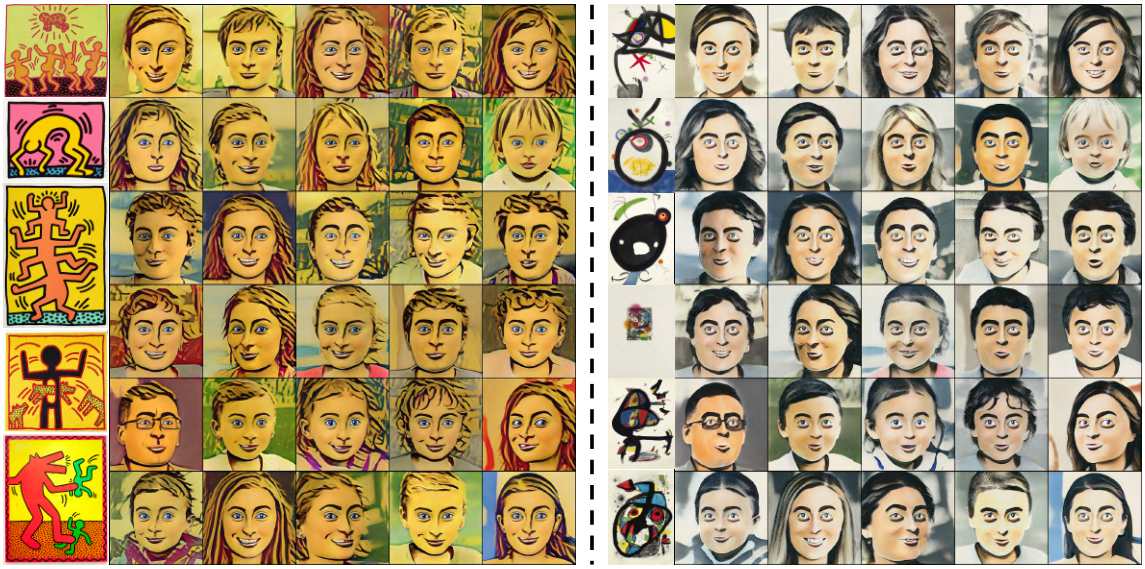


Figure 4.9: Image synthesis results on human face

### 4.4.4 Additional Image Synthesis Results

In order to further explore the effect of the few-shot image guidance, we utilize our proposed data factory to synthesize other objects, human face, adapted by the text guidance "clipart" in combination with few-shot images of different artistic styles. The qualitative results are shown in Fig. 4.9. The image synthesis is guided by the text "clipart" and the few-shot image samples from "Keith Harring" and "Joan Miro" style paintings. The first column of the left part and the right part are examples of the "Keith Harring" and "Joan Miro" painting styles. It proves that our proposed data factory effectively synthesizes the target-domain-like images under the text and the few-shot image guidance, and effectively reflects the difference under one general style category. Pre-acquired text knowledge can be ambiguous and general under a lot of application scenarios. Compared to previous works based purely on text guidance, our approach enables more detailed and explicit control by combining extra guidance images.

Moreover, to show the possible application of our method to real scenes, we conduct experiments adapting original images of the car category to reflect different weathers, *i.e.*foggy, rainy and snowy. Fig 4.10 shows the adapted synthesized images under different weather conditions. For the adaptation of each setting, text description from "Sunny" to the specific weather together with 5 style images of street scenes under such weather condition are provided for guidance. As shown in the figure, our image synthesis adaptation

approach resembles natural scenes for all three settings.


Figure 4.10: Image synthesis results on car, with few-shot image and text guidance of "foggy", "rainy", and "snowy"

### 4.4.5 Label Synthesis for Multiple Classes

In order to prove the effectiveness of our label synthesis branch to be extended to multi-class scenarios, we implement an additional experiment on an indoor scene with multiple objects to be detected in one image. We take the pretrained StyleGAN2 model on the bedroom dataset [35], and manually annotate 25 training samples with 4 categories: bed, lamp, table, and window. The label synthesis branch is trained under the same settings as the benchmark experiments. Visualized detection results are shown in Fig. 4.11. Although we have not fine-tuned the training process to optimize the model performance, it generates valid predictions, indicating our approach's possibility to generalize to multi-category applications.

### 4.4.6 Pseudo Label *vs.* Our Label Synthesis

Under the cross-domain experiments setting, an alternative way to our label synthesis through the efficient labeled data factory is to apply the source domain pretrained object detection model on our synthesized images to generate the pseudo-label. In Table 4.6, the pseudo-label and the label synthesis with our efficient labeled data factory ways for

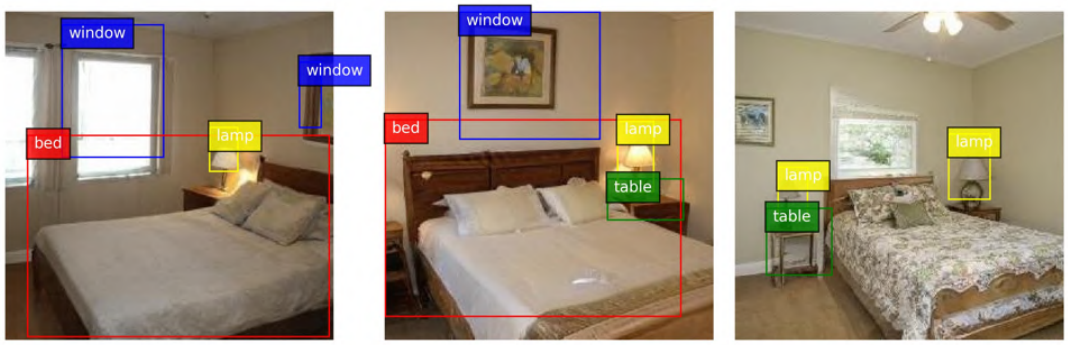


Figure 4.11: Qualitative results of synthesized labeled data of bedroom with four classes: bed, lamp, table, window

label generation are compared. It is shown that our synthesized label with the data factory performs better than the pseudo-label for the SF-FSDA problem. It is because the pseudo-label is noisy and of low quality, resulting from the difference between the source domain and the synthesized images, and the source-free condition. In contrast, our efficient labeled data factory synthesizes the annotation with the help of the image synthesis and few-shot manual annotations, bringing high-quality automatic annotations.

| | Pseudo-Label | Our Label Synthesis |
|---|---|---|
| Cat | 25.75 | **32.50** |
| Car | 53.52 | **55.67** |

Table 4.6: Comparison of label generation ways, pseudo label *vs.* our data factory, PASCAL VOC→ Clipart

# Chapter 5

# Discussion

In this chapter, we explore other image style adaptation approaches as prospective alternative solutions for the image synthesis branch. The possible future work is also discussed.

## 5.1 Image Style Adaptation with GAN Inversion

In this section, we present other works we explore during this thesis, which leverage further possibilities of image style adaptation. In our work, we focus on the direction of fine-tuning a pretrained GAN model with style supervision. As discussed in Sec. 2.2, a lot of works take advantage of encoding in the latent space via GAN inversion. The general drawback in this field lies in the requisition of an extra encoder that also requires considerable training effort as the GAN model. An encoder for GAN inversion takes an image and encodes it into the GAN latent space. The generated latent code should allow for high-quality reconstruction of the original image when passed to the generator of the GAN model. Thus, given a pretrained encoder for image inversion, image manipulation could be done by first generating the encoded latent vector of this image, then manipulating the content/style by editing the latent code. Here we present our trials in this field and discuss the potentials as well as their own limitations under our proposed SF-FSDA problem.

### 5.1.1 Latent Code Editing

One way of achieving the desired manipulation is to explore significant editing directions for different properties, based on experimentally examined results [79], or decomposition methods [24, 67]. Instead, it is also possible to directly optimize the inverted latent code by providing extra training guidance. [56], as the very initial work of incorporating CLIP

with StyleGAN for image style manipulation, changes the style of an image by optimizing its inverted latent code. In this work, they propose three ways for image manipulation. The first is to directly optimize the inverted latent code of the image to be modified, to resemble the guidance text in the embedded CLIP space. The second is to train a latent mapper to learn the edit vector that could to applied to inverted latent codes of all images under the same category, *e.g.*learning a smiling face for human portraits. The third method is to learn a global editing direction in the latent code space, in which the image changes the same with the text guidance. In our efficient labeled data factory based approach, we introduce the text guidance by enforcing the generated image to change along the text guidance direction in the CLIP embedded space, similar to the first approach. We also explore the second and the third approach, training a model to find meaningful optimization in the latent feature space instead of fine-tuning the GAN network. The global direction approach requires training image pairs indicating the same changing direction as the text guidance to find the layer-wise image editing vector in the GAN style space, which are hard to be acquired, thus does not fit under our problem setting. For the latent mapper approach, as for our SF-FSDA setting with no access to the source images, we take synthesized images generated by the pretrained GAN model as our training images, and therefore, no additional encoder is required here. According to our experiment results, training without real images highly defects the performance due to poor representation in the low-density latent space area, which makes this approach not an optimal solution for our problem.

### 5.1.2 Style Mixing

Instead of traversing latent code editing directions or editing vectors, style mixing [63] could be achieved by inverting the style image and mixing its latent code with the inverted vector of a real image or a randomly sampled latent code to generate the image in desired style or content. Fig 5.1 shows our experiments on the cat category, with a pSp encoder [63] pretrained on the AFHQ cat dataset [11] and StyleGAN2 model pretrained on the cat category. Fig 5.1(a) shows the style image for guidance and its reconstruction result with the inverted latent code obtained by the pSp encoder. (b) shows the original images generated by specific sampled latent codes. (c)-(f) present the images generated by replacing certain layers of the randomly sampled latent codes with the inverted style vector of the guidance image in the $\mathcal{W}+$ space. As shown from the figure, different levels of style, *e.g.*pose, color, facial features could be controlled to resemble the provided style image by swapping certain layers due to the well-disentangled nature in the StyleGAN $\mathcal{W}+$ latent space. Moreover, this could be extended to the multi-modal scenario as shown in Fig 5.2 where the same swapping is applied to the same sampled latent vectors and then forwarded to the adapted generator with style in another domain (clipart in this

case). Although it could be applied to some simple multi-modal conditions, *e.g.*moderate color and/or pattern variation in artistic styles, such approach still does not provide a valid solution to the SF-FSDA problem under our setting, due to the bias of unrelated content of the source and target domains (see Fig 3.4 in Sec. 3.3.1). The reconstruction results of latent codes inverted by e4e [68] and pSp [63] encoders pretrained on the FFHQ dataset are shown in Fig. 5.3. (a) is a human portrait within the pretrained domain. (b) and (c) are cartoon portraits that share similar contents with the original domain. (d) shows an image from a more distant domain, while the content of (e) is fully unrelated to the original domain. The encoder is only capable of inverting images within the pretrained category. Passing style images of unrelated content does not generate meaningful latent codes, thus not being able to be mixed for style adaptation. Still, it leverages the possibility of further controlling the synthesized images of our data factory, as shown in Fig 5.2.


Figure 5.1: Style mixing of the pretrained generator in the real cat domain with a real cat image

## 5.2 Future Work

In this section, we present possible future works to improve and extend our current approach.

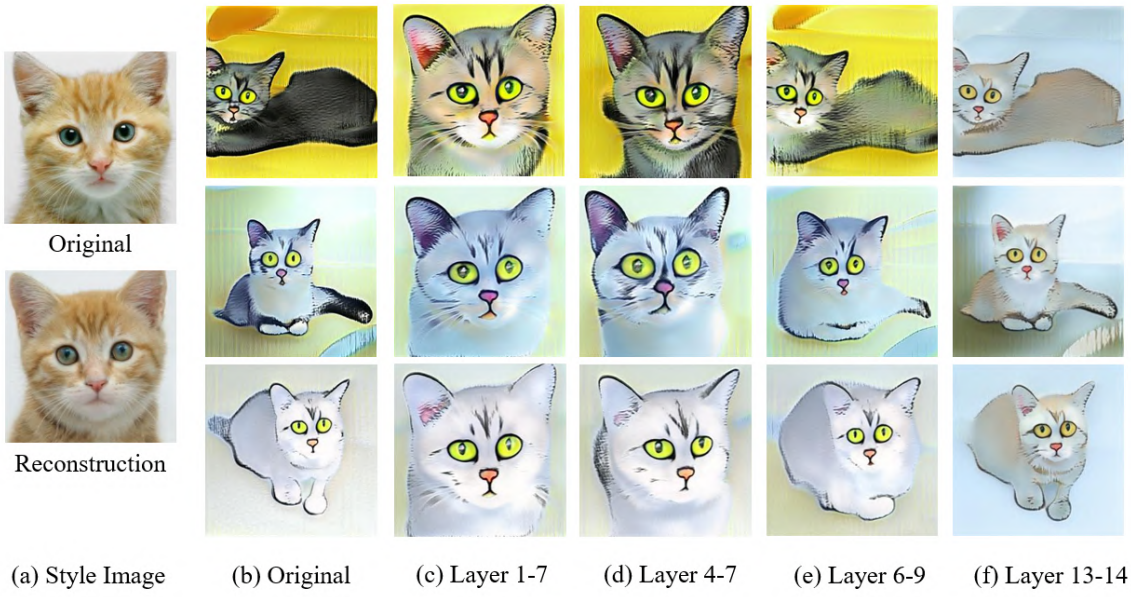


Figure 5.2: Style mixing of the pretrained generator in the clipart style cat domain with a real cat image

**More Controllable Image Synthesis Adaptation** As discussed in Sec. 5.1, more control of both content and style of the synthesized images could be achieved via incorporating guidance style vectors. Introducing such style vectors not only leverages a more controllable process, but also helps generate higher-quality images with better represented latent code inverted from real images. Besides, we control our training process by introducing the freezing strategy, as discussed in Sec.3.3.3. We intuitively freeze the shallow layers for overfitting prevention. As discussed in [80], layer-wise control and semantic relevance of StyleGAN still remain valid under the standard fine-tuning procedure of GANs. Such nature could be further explored under our condition with limited guidance and help better regularize the adaptation process.

**Improving Label Synthesis Network** Currently, we follow most default settings in the original work [87]. Theoretically, our model could still benefit from further fine-tuning, both for our benchmarks settings and the multi-class detection case discussed in Sec. 4.4.5. Besides, [86] demonstrates promising pixel-wise prediction for the semantic segmentation task. While in our case, we obtain valid results in object detection tasks localizing object instances. We see the potential of extending the prediction task utilizing StyleGAN semantic feature maps to panoptic segmentation [38], which unifies the semantic segmentation and instance segmentation problems to generate object-wise segmentation

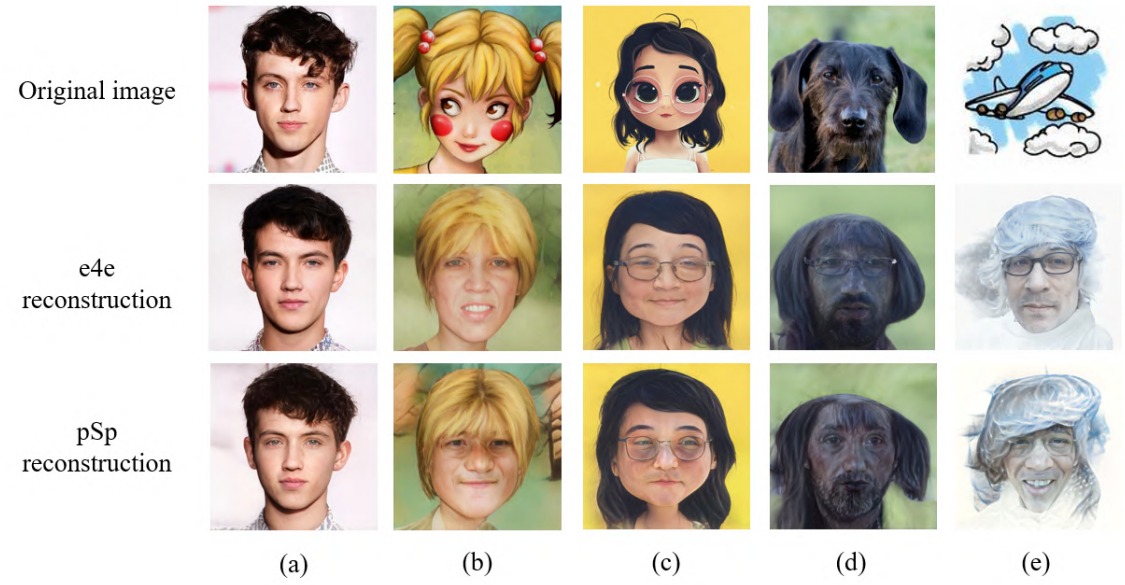


Figure 5.3: Reconstruction results of e4e and pSp encoder with images within/outside the pretrained domain

with categories on the pixel level. Some recent works [83, 10] solve this problem via bottom-up approaches, starting with semantic segmentation predictions and then grouping the segmentation results into clusters to get the instance masks. The grouping operation is typically done by predicting object centers and then regressing around the centers to get the instance segmentation for each object. Combining the work of [86] and ours leverages it as a possible future work to achieve panoptic segmentation for our label synthesis branch, and thus to benefit a wider range of cross-domain downstream tasks.

# Chapter 6

# Conclusion

We propose and tackle the SF-FSDA problem, which studies the domain adaptive object detection problem under *source-free* and *few-shot* conditions. In order to overcome the problem, we present a new efficient labeled data factory based method, which can synthesize the infinite target-domain-like images and corresponding annotations without relying on the source domain. The image synthesis branch is guided by the few-shot image samples and text from the target domain, and the image annotation branch only requires the minimum human effort (*i.e.*, few-shot manual labels) to generalize the label to the rest of the synthesized images. The synthesized target-domain-like images and annotations are further utilized to fine-tune the source domain pretrained object detection model, realizing robust object detection. The proposed approach is validated in various settings and surpasses other state-of-the-art methods, demonstrating its effectiveness for the SF-FSDA problem.

# Bibliography


[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.

[3] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

[4] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *CVPR*, 2020.

[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[8] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *CVPR*, 2016.

[9] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

[10] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[12] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, 2020.

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

[14] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021.

[15] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *CVPR*, 2020.

[16] Chris Donahue, Zachary C Lipton, Akshay Balsubramani, and Julian McAuley. Semantically decomposing the latent spaces of generative adversarial networks. *arXiv preprint arXiv:1705.07904*, 2017.

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[18] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.

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

[20] Hongchang Gao, Jian Pei, and Heng Huang. Progan: Network embedding via proximity generative adversarial network. In *SIGKDD*, 2019.

[21] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021.

[22] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

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

[24] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

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

[26] Rani Horev. Explained: A style-based generator architecture for gans-generating and tuning realistic artificial faces, 2018.

[27] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, 2020.

[28] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, 2020.

[29] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[30] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018.

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[32] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.

[33] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.

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

[35] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

[36] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, 2019.

[37] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, 2019.

[38] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[39] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, 2021.

[40] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *CVPR*, 2020.

[41] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, Rahul M V, and R. Venkatesh Babu. Towards inheritable models for open-set domain adaptation. In *CVPR*, 2020.

[42] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *CVPR*, 2020.

[43] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021.

[44] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *AAAI*, 2020.

[45] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[46] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *NeurIPS*, 2021.

[47] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019.

[48] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

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

[50] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *ICLR*, 2019.

[51] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in neural information processing systems*, 29, 2016.

[52] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *NeurIPS*, 2017.

[53] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. In *NeurIPS*, 2021.

[54] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, 2021.

[55] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020.

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

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

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

[59] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *ICCV*, 2021.

[60] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[62] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *ICCV*, 2021.

[63] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[64] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *CVPR*, 2019.

[65] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In *ECCV*, 2020.

[66] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019.

[67] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021.

[68] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[69] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021.

[70] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.

[71] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[72] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[73] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A Sindagi, and Vishal M Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, 2021.

[74] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

[75] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14050–14060, 2021.

[76] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *CVPR*, 2019.

[77] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *ICCV*, 2017.

[78] Qilong Wu, Xiangyu Yue, and Alberto Sangiovanni-Vincentelli. Domain-agnostic test-time adaptation by prototypical training with auxiliary data. In *NeurIPS Workshop*, 2021.

[79] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021.

[80] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021.

[81] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2021.

[82] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *ICCV*, 2021.

[83] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*, 2019.

[84] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019.

[85] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 41(8):1947–1962, 2018.

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

[87] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, 2019.

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

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

[90] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398*, 2021.

[91] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036*, 2020.