SF-FSDA: Source-Free Few-Shot Domain Adaptive Object Detection with Efficient Labeled Data Factory

Han Sun
EPFL
Switzerland
han.sun@epfl.ch

Rui Gong
ETH Zurich
Switzerland
gongr@vision.ee.ethz.ch

Konrad Schindler
ETH Zurich
Switzerland
schindler@ethz.ch

Luc Van Gool
ETH Zurich
Switzerland
vangool@vision.ee.ethz.ch

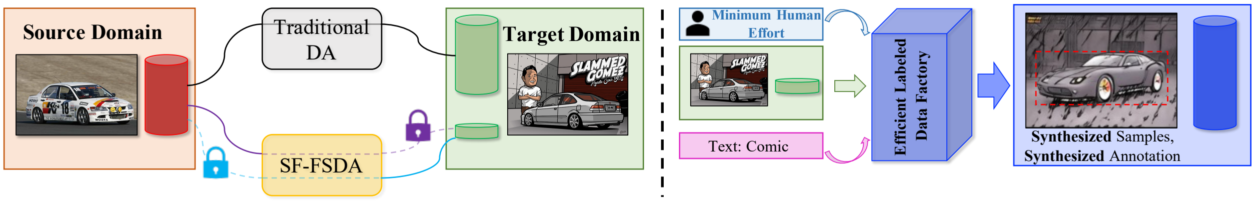
Abstract

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 the access to the source domain data for adaptation, and the availability of sufficient data on the target domain. However, these assumptions may not hold due to data privacy and rare data collection. In this paper, 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 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 the robust object detection under SF-FSDA. The experiments on different settings showcase that our proposed approach outperforms other state-of-the-art methods on SF-FSDA problem. Our codes and models will be made publicly available.

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 Ren et al. (2015); Liu et al. (2016); Redmon et al. (2016); Carion et al. (2020). 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 a different distribution with 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 Chen et al. (2018); Saito et al. (2019); Hsu et al. (2020a); Khodabandeh et al. (2019); VS et al. (2021); Ramamonjison et al. (2021), which aims to transfer the knowledge learned from the labeled source domain to the unlabeled target domain to train the 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 Hoffman et al. (2018); Inoue et al. (2018); Bhattacharjee et al. (2020) and/or feature space Chen et al. (2018); Deng et al. (2021)) through the typical techniques of adversarial learning Rezaeianaran et al. (2021); Saito et al. (2019), pseudo-label RoyChowdhury et al. (2019); Kim et al. (2019); Munir et al. (2021), and image translation Hsu et al. (2020b);



(a) SF-FSDA vs Traditional DA

(b) Efficient Labeled Data Factory

Figure 1: (a) Traditional domain adaptive object detection (DA) problem *vs*. our proposed SF-FSDA problem. Our SF-FSDA problem considers the *few-shot* and *source-free* conditions. (b) Our efficient labeled data factory can synthesize abundant target domain like image samples and corresponding bounding box and category annotation automatically, by providing the few-shot samples guidance, text guidance, and minimum human effort (*i.e*., few-shot manual label). The efficient labeled data factory is initialized with publicly available pretrained GAN model weights, and does not rely on the access to the source domain images and/or other public image dataset. The height of the cylinder represents the number of image samples.

Inoue et al. (2018). 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. Even though some more recent works touch the domain adaptive object detection problem under the source-free condition Li et al. (2020b) or the few-shot condition Wang et al. (2019), none of them consider the problem when both the source-free and few-shot conditions exist at the same time. Instead, this work studies 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. 1a).

The current available domain adaptive object detection approaches can tackle isolated one of the two conditions in SF-FSDA, but cannot deal with the two conditions at the same time. More specifically, pseudo-label based techniques are popularly utilized in source-free conditions Li et al. (2020b); Kundu et al. (2021), but are not capable of handling the few-shot condition, since it relies on enough samples to reduce the pseudo-supervision noise brought by the domain gap. In contrast, adversarial learning Motiian et al. (2017); Ganin & Lempitsky (2015) and image translation Luo et al. (2020); Ma et al. (2019) based methods can operate under few-shot conditions, but require access to the source domain.

In order to address the challenging SF-FSDA problem, we propose an efficient labeled data factory based method (see Fig. 1b), 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. Compared to the existing image translation based approach Zhu et al. (2017); Park et al. (2020); Huang et al. (2018); Liu et al. (2019); Isola et al. (2017), our proposed data factory based method does not require the availability to the source domain (*source-free condition*), and effectively exploits the few-shot image (*few-shot condition*) and text knowledge from the target domain for the image synthesis. The text knowledge provides the general guidance, *e.g*., what the "comic" style images look like, while the few-shot images offer the specific guidance, *e.g*., how the "comic" images on the target domain are like. The text knowledge can also prevent the overfitting effect brought by the few-shot images. Thus, the text and few-shot images guidances promote each other to synthesize the more target-domain like images. Besides, the data factory based method renders the object bounding box and category annotation together with and guided by the image synthesis process, only requiring the few-shot manual annotation. Our data factory model is initialized with publicly available GAN pretrained weights, which are irrelevant to both the source and target domain. In this way, the efficient labeled data factory synthesizes the target domain on both the *image level* and *label level*. Alternatively, applying the source domain trained object detection model on our synthesized image can generate the pseudo-label, which however is noisy and low-quality due to the domain gap. Instead, our efficient labeled data factory can synthesize the detection label without relying on the source trained object detection model, generating higher quality annotation and easing the downstream domain adaptive object detection. Then the source domain trained object detection model is fine-tuned on our synthesized images and annotations to train the final object detection model.

In a nutshell, the key contributions of this paper are three-fold. **(1)** We propose the domain-adaptive object detection problem under the *source-free* and *few-shot* conditions, SF-FSDA, where there are only a few samples available on the target domain, and only the source-pretrained model is accessible for the adaptation to the target domain. **(2)** We develop the efficient labeled data factory based approach to SF-FSDA problem, where the efficient labeled data factory can automatically synthesize a number of the target domain like images and corresponding object detection labels, providing the text guidance, few-shot sample guidance and minimum human annotation. **(3)** The experimental results on different benchmarks prove the effectiveness of the proposed method for the SF-FSDA problem, serving as the strong baseline for further research.

2 Related Work

2.1 Image Synthesis.

Recently, generative adversarial networks (GANs) Goodfellow et al. (2014) 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 Brock et al. (2018); Gao et al. (2019); Zhang et al. (2018) are able to generate high-quality realistic images of diverse categories. Recent style-based generators Karras et al. (2019; 2020b;a; 2021) produce impressive results and allow for style control via mapping noise vectors to a higher-dimensional semantic space, which inspires several extensions such as image manipulation Patashnik et al. (2021b); Gal et al. (2021); Zhu et al. (2021), image editing Ling et al. (2021); Bau et al. (2021), and dataset synthesis Zhang et al. (2021). Besides, various types of potential guidance Li et al. (2020a); Ling et al. (2021); Dhamo et al. (2020); Collins et al. (2020); Patashnik et al. (2021a) are utilized for controlling the image synthesis process, among which the most relevant to our work are text guidance and few-shot image guidance. The text guidance is typically provided by the large-scale vision-language pretrained model Radford et al. (2021), by mapping the images and the text to the joint embedding space. More recent works Patashnik et al. (2021b); Gal et al. (2021) then introduce the image and text consistency in the embedding space to regularize image synthesis. The few-shot image guided image synthesis works Saito et al. (2020); Liu et al. (2019); Ojha et al. (2021) aim to prevent image generation overfitting when only limited image samples are available. 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 (see ablation study in Sec. 4.2). The label synthesis of our data factory is related to the work of Zhang et al. (2021), which, however, only focuses on the dense prediction task, *e.g.*, semantic segmentation, and does not consider the domain adaptation problem. Instead, 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.

2.2 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 Tzeng et al. (2017); Gong et al. (2012); Ganin & Lempitsky (2015), semantic segmentation Tsai et al. (2018); Vu et al. (2019); Tranheden et al. (2021), and object detection Chen et al. (2018); Khodabandeh et al. (2019); VS et al. (2021). 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 Li et al. (2020b) or few-shot Wang et al. (2019) domain adaptive object detection problem, respectively. More specifically, Li et al. (2020b) tackles the source-free Kundu et al. (2020a); Yang et al. (2021); Kundu et al. (2020b) domain adaptive object detection problem with the pseudo-label based technique. And Wang et al. (2019) studies the few-shot Motiian et al. (2017) 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.

2.3 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 image is not available for training, Yue et al. (2019) randomizes the style of the source domain images utilizing the images from the public dataset ImageNet Deng et al. (2009), to improve the generalization ability of the semantic segmentation model trained on the source domain. Wu et al. (2021) 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 Karras et al. (2020b) as the auxiliary knowledge, which is more flexible and renders unlimited and unified image and label synthesis.

3 Method

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 $\mathbf{x}_s^i, \mathbf{y}_s^i$ 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.

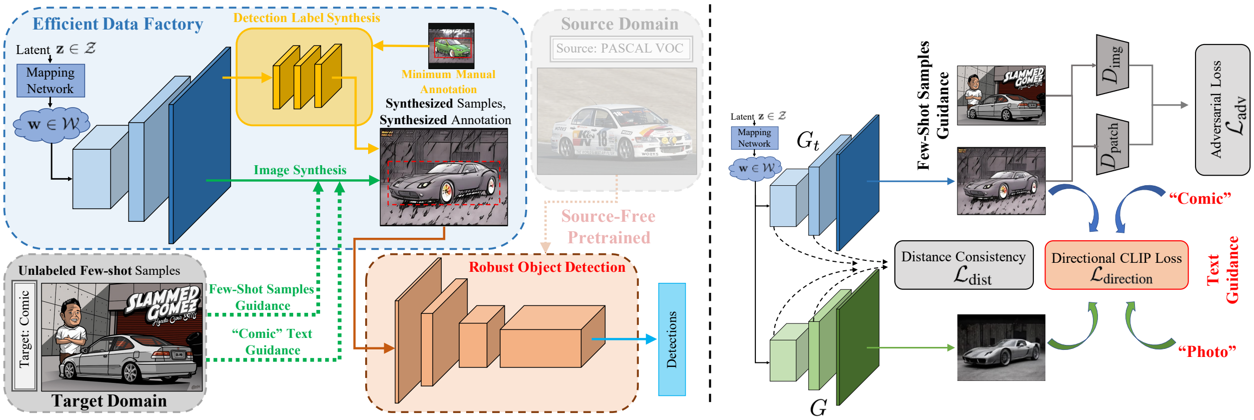
Technical Challenges. Compared to the traditional domain adaptive object detection problem, our proposed SF-FSDA problem introduces more challenging *source-free* and *few-shot* conditions. Previous techniques for domain adaptive object detection highly rely on adversarial feature learning Chen et al. (2018), image-to-image translation Inoue et al. (2018), and pseudo-label based self-training RoyChowdhury et al. (2019). On the one hand, the challenge brought by *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 *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.

Motivation. As discussed in the aforementioned technical challenges, the *source-free* and *few-shot* conditions make it difficult to adapt guided by the object detection task. Thus, we aim to firstly adapt on the image level, *i.e.*, synthesize the target domain like image. However, different from the previous image translation methods that rely on the access to the source domain and the target domain at the same time, the adaptation, guided by the few-shot samples and text from the target domain, on the publicly available trained GAN model provides more flexibility without accessing the source domain data. Moreover, in order to provide reliable guidance for the downstream object detection task, the method for synthesizing the corresponding object detection label 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, as shown in Fig. 2. 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, the object detection model is trained on the source domain. Then in the ii) image and label synthesis stage, the efficient labeled data factory is driven by the few-shot image sample and text guidance from the target domain to synthesize the image with the image synthesis branch, and the label synthesis branch automatically synthesizes the corresponding object detection label by only providing the few-shot manual annotation. In the iii) target-adaptation stage, the synthesized image and corresponding label synthesized in stage ii) are exploited to fine-tune the source pretrained object detection model in stage i).

Image Synthesis with Few-Shot Image Guidance. 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 $\mathcal{T}$. Following Ojha et al. (2021), the distance consistency regularization, $\mathcal{L}_{dist}$, is utilized to preserve the original content and diversity of the image samples, and the anchor-based relaxed realism is adopted to further prevent the overfitting


Efficient Data Factory
Latent z ∈ 𝒵
Mapping Network
w ∈ 𝒲
Detection Label Synthesis
Minimum Manual Annotation
Synthesized Samples, Synthesized Annotation
Image Synthesis
Source Domain
Source: PASCAL VOC
Source-Free Pretrained
Unlabeled Few-shot Samples
Target: Comic
Target Domain
Few-Shot Samples Guidance
"Comic" Text Guidance
Robust Object Detection
Detections
Latent z ∈ 𝒵
Mapping Network
w ∈ 𝒲
G_t
G
Few-Shot Samples Guidance
D_{img}
D_{patch}
Adversarial Loss $\mathcal{L}_{\text{adv}}$
"Comic"
Distance Consistency $\mathcal{L}_{\text{dist}}$
Directional CLIP Loss $\mathcal{L}_{\text{direction}}$
Text Guidance
"Photo"

(a) Method Overview

(b) Image Synthesis Overview

Figure 2: (a) Method overview. The well labeled PASCAL VOC dataset is taken 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* condition, *i.e.*, only the source pretrained model and few-shot target domain samples are available for the adaptation to the target domain. The efficient data factory module is composed of the image synthesis branch and the detection label synthesis branch. The image synthesis branch is guided by the few-shot samples and text from the target domain, to synthesize the target domain-like image. The detection label synthesis branch aims to automatically synthesize the bounding box and category label, with the help of minimum human effort, *i.e.*, few-shot manual annotation. (b) Image synthesis overview. The image synthesis branch is driven by the few-shot image samples guidance and the text guidance.

to the few-shot image samples. In more detail, the distance consistency regularization samples a batch of $N+1$ noise vectors and uses their pairwise similarities in feature space to construct an N-way probability distributions for each image. The N-way probability of the image generated by the i-th noise vector is given by:

$$\begin{aligned} y_i^m &= \text{Softmax}\left(\{\text{sim}\left(G^m\left(z_i\right), G^m\left(z_j\right)\right)\}_{\forall i \neq j}\right) \\ y_i^{t,m} &= \text{Softmax}\left(\{\text{sim}\left(G_t^m\left(z_i\right), G_t^m\left(z_j\right)\right)\}_{\forall i \neq j}\right), \end{aligned} \tag{1}$$

where similarity denotes the cosine similarity of activations at the m^{th} layer between generator G and G_t. The probability distributions converted from the similarities of the adapted model and the given publicly pretrained model are encouraged to be uniform by computing KL-divergence across the intermediate layers. With the anchor-based relaxed realism, a dual-discriminator training approach is deployed to prevent overfitting to the few-shot samples. 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.

We observe minor collapse in style with the training strategy in Ojha et al. (2021) under our setting, *i.e.*, the same color or pattern before successful adaptation. Different from Ojha et al. (2021), we relax the distance consistency regularization during different training phrases to allow a reasonable extent of object shape adaptation while still keeping the original image content. In the initial training phases, we compute the distance consistency only of the deep layers (*i.e.*, after 6th layer) of the generator. After training for certain epochs, we adapt the training strategy and only compare consistency on the shallow layers (*i.e.*, before 10th layer) for detailed style adaptation and to preserve the content.

The objective of image synthesis with few-shot image guidance is,

$$G_t^* = \arg\min_{G_t} \max_{D_{\text{img}}, D_{\text{patch}}} \mathcal{L}_{\text{adv}}\left(G_t, D_{\text{img}}, D_{\text{patch}}\right) + \lambda_1 \mathcal{L}_{\text{dist}}\left(G_t, G\right), \tag{2}$$

where $\mathcal{L}_{\text{adv}}$ represents the adversarial loss, and λ_1 is the hyper-parameter to balance the adversarial loss and the distance consistency regularization loss.

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." 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 contrastive language-image pre-training

(CLIP) models Radford et al. (2021). 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 Gal et al. (2021). Original and target texts are both self-defined to provide the desired shifting guidance. In order to obtain the image shifting direction during training, 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 text guidance and images can be expressed by,

$$\begin{aligned} \Delta T &= E_{text}\left(T_{\text{target}}\right) - E_{text}\left(T\right) \\ \Delta I &= E_{img}\left(G_t(\mathbf{z})\right) - E_{img}\left(G(\mathbf{z})\right), \end{aligned} \tag{3}$$

where E_{text} and E_{img} denote CLIP text and image encoders, respectively. T and T_{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,

$$\mathcal{L}_{\text{direction}}(G, G_t, T, T_{\text{target}}) = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|}. \tag{4}$$

Combined with the few-shot image guidance training objective in Eq.(2), our final training objective with both the few-shot image guidance and the text guidance can be derived as,

$$\begin{aligned} G_t^* = \arg\min_{G_t} \max_{D_{\text{img}}, D_{\text{patch}}} & \mathcal{L}_{\text{adv}}\left(G_t, D_{\text{img}}, D_{\text{patch}}\right) \\ & + \lambda_1 \mathcal{L}_{\text{dist}}\left(G_t, G\right) \\ & + \lambda_2 \mathcal{L}_{\text{direction}}\left(G_t, G, T, T_{\text{target}}\right), \end{aligned} \tag{5}$$

where λ_2 is the hyper-parameter to balance the text guidance and other terms.

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. More specifically, the shallow layers of the original generator are frozen, while shallow layers of the discriminator are also frozen accordingly to further ensure a stable training process. The training strategy is simple yet effective in preventing overfitting under few-shot conditions.

Label Synthesis. We can now get unlimited target-like samples with a successfully adapted image synthesis branch. Previous research has proved that StyleGAN2 Karras et al. (2020b) learns a well-disentangled semantic latent space, where each channel controls some 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 and 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 $\{G_t(z_i)\}_{i=1}^{N_a}$. Here N_a denotes the number of manual annotations required to train the object detection task. 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 concatenate the intermediate convolutional feature maps generated by the latent codes as our encoded features for the matching images. A prediction head is built on these extracted features and trained for the object detection task.

Inspired by Zhou et al. (2019), 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 Lin et al. (2018),

$$\mathcal{L}_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1-\hat{\mathbf{y}}_{xyc})^{\alpha} \log\left(\hat{\mathbf{y}}_{xyc}\right) & \text{if } \mathbf{y}_{xyc} = 1 \\ (1-\mathbf{y}_{xyc})^{\beta} \left(\hat{\mathbf{y}}_{xyc}\right)^{\alpha} \log\left(1-\hat{\mathbf{y}}_{xyc}\right) & \text{otherwise,} \end{cases} \tag{6}$$

where α and β 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,

$$L_{off} = \frac{1}{N}\sum_{p}\left|\hat{\mathbf{o}}_{\tilde{p}} - \left(\frac{p}{r} - \tilde{p}\right)\right|$$

$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{\mathbf{s}}_{p_k} - \mathbf{s}_k\right|. \tag{7}$$

A two-layer convolutional head is built as the label synthesis branch for predicting $\hat{\mathbf{y}}$, $\hat{\mathbf{o}}$, and $\hat{\mathbf{s}}$ each and trained with a weighted sum of loss terms for these tasks,

$$\mathcal{L}_{det} = \mathcal{L}_k + \lambda_{off}\mathcal{L}_{off} + \lambda_{size}\mathcal{L}_{size}, \tag{8}$$

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

4 Experiment

4.1 Experimental Setup

In-Domain Experiments. In order to verify the validity of our proposed efficient labeled data factory for automatically producing the images and 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 on 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. 3c) synthesized by our data factory, and tested on the PASCAL VOC datset.

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) → Clipart and Comic (target), respectively.

Dataset. *PASCAL VOC*: PASCAL VOC 2007 & 2012 datasets Everingham et al. (2010) contain natural objects with manually labeled bounding box and category annotations. 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. *Clipart1k*: Clipart1k dataset Inoue et al. (2018) 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. *Comic2k*: Comic2k dataset Inoue et al. (2018) 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.

Training Details. *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 Karras et al. (2020b). *Label Synthesis*: As the minimum human effort, we manually label 10 synthesized images. *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) Liu et al. (2016) model. We synthesize 200 and 250 samples in total in the in-domain and cross-domain experiments, respectively.

Baseline Setup. In Table 1 and Table 3, the "Few-Shot FT" represents that the object detection model is fine-tuned on the few-shot manually-labeled images from our data factory. In Table 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 1 is reached by training the object detection model on the training set of PASCAL VOC. Oracle performance in Table 3 is obtained by Inoue et al. (2018) for the traditional domain adaptive object detection, which is not few-shot or source-free.

Table 1: In-domain experiments on PASCAL VOC. The results are reported on average precision (AP).

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

Table 2: Comparison of label generation ways, pseudo label *vs.* our data factory, PASCAL VOC→ Clipart.

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

4.2 Experimental Results

Through the quantitative and qualitative in-domain and SF-FSDA cross-domain experimental results, we show that our proposed model effectively synthesizes the image samples and the corresponding object bounding box and category labels, and mitigates the source and target domain gap through the guidance of text and few-shot target examples.

In-Domain Experiments. As shown in Table 1 and Fig. 3c, 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 box 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.

SF-FSDA Cross-Domain Experiments. In Table 3 and Fig. 3d-g, the quantitative and qualitative results are shown on the benchmark, PASCAL VOC→ Clipart, Comic, respectively. Taking the PASCAL VOC→ Clipart benchmark as the example, compared with the pure source baseline, all of 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 Zhu et al. (2017), MUNIT Huang et al. (2018), and CUT Park et al. (2020). It proves the advantage of our method for synthesizing the target-domain like image, with the guidance of both the few-shot samples and the text knowledge. Moreover, compared with the few-shot manual annotations, the automatically synthesized annotations can further improve the performance from 30.94%, 52.97% to 32.50%, 55.67%. It verifies the effectiveness of the automatically generated images and annotations for the SF-FSDA problem. Similarly, on the PASCAL VOC→ Comic benchmark, the effectiveness of our proposed method for SF-FSDA is also validated.

Ablation Study. 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 different types of guidance, we compare the performance of different ablations of the full model. From the quantitative comparison in Table 4a, it is shown that both the few-shot samples and text guidance contribute to the final image synthesis results. From the qualitative results shown in Fig. 3h-j, taking the "comic" style as the example, the text guidance provides the general knowledge on what the "comic" images look like, while the few-shot images guidance indicates how the "comic" 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. Moreover, the ablation study on the freezing strategy during training is conducted, which is shown in Fig. 3k-l and Table 4b. It is shown that the freezing

Table 3: SF-FSDA cross-domain experiments.

Classes	Method						
	Source	Few-Shot FT	CycleGAN	MUNIT	CUT	Ours	Oracle
PASCAL VOC → Clipart							
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
PASCAL VOC → Comic							
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: Ablation study, PASCAL VOC → Clipart (Cat).

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

(a) Ablation study for the text and few-shot images guidance from the target domain, measured with AP performance on Clipart.

Ablations	
w/o freezing	w freezing
0.64	**0.68**

(b) Ablation study for freezing strategy during image synthesis training, measured with the LPIPS distance Ojha et al. (2021)(↑).

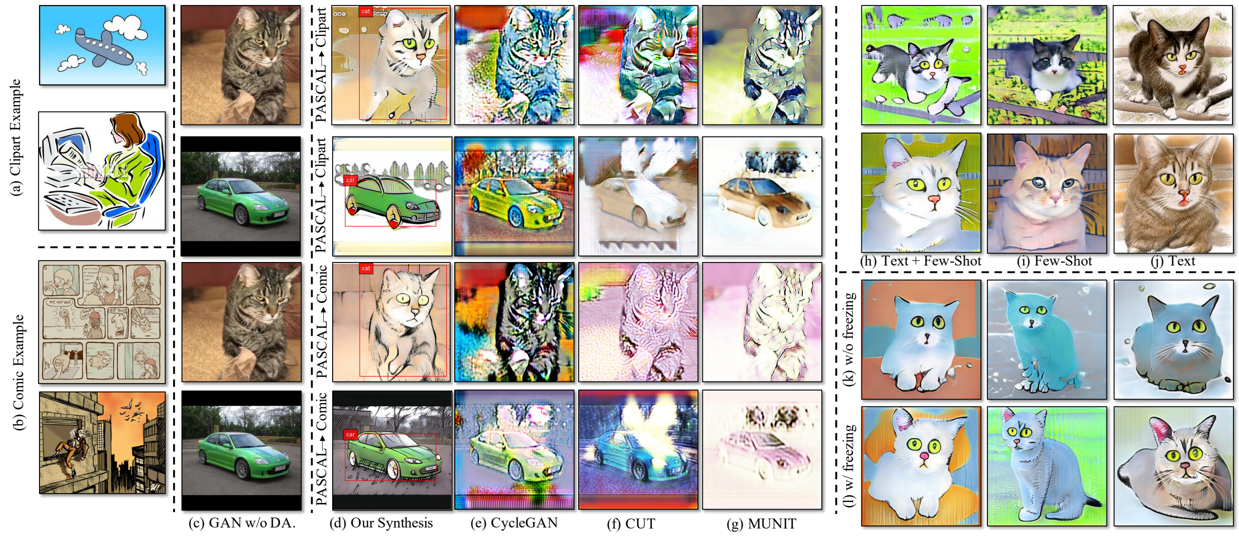


Figure 3: (a)-(b) are the exemplar images from the Clipart1k and Comic2k dataset. (c) are the synthesized images from the publicly available pretrained GAN weights, without conducting domain adaptation and used in Table 1. It is notable that our data factory does not have the requirement of on which style images the GAN model is pretrained, and we just adopt the publicly available pretrained weights provided in Karras et al. (2020b). (d) are the synthesized image and annotations from our proposed data factory in Table 3. (e)-(g) are the results generated by other image translation methods in Table 3. (h)-(j) are the ablations for guidance as in Table 4a. (k)-(l) are the ablations w/ and w/o freezing strategy as in Table 4b.

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.

Source Domain Pretraining. To investigate the impact of a source domain pretrained object detection model on SF-FSDA, we conducted a comparison of the SF-FSDA performance with and without the pretrained model for the cat of PASCAL VOC→Clipart dataset. For the experiment with pretrained model, the pretrained object detection model on source domain is fine-tuned with our generated images and labels on the target domain as done in Sec. 3.2. For the experiment without pretrained model, the object detection model is trained from scratch with our generated images and labels on the target domain. Our results show that the SF-FSDA performance significantly improves with the use of a pretrained model from a large-scale source domain, yielding 32.50% accuracy with source pretrained compared to 17.74% without source pretrained. This suggests that leveraging a pretrained model from a relevant large-scale source domain can provide substantial benefits for the success of SF-FSDA.

Number of Synthesized Image Samples Study. In order to 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. It is shown that the object detection performance improves as more images and annotations are synthesized.

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 2, the pseudo-label and the label synthesis with our efficient labeled data factory ways for 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.

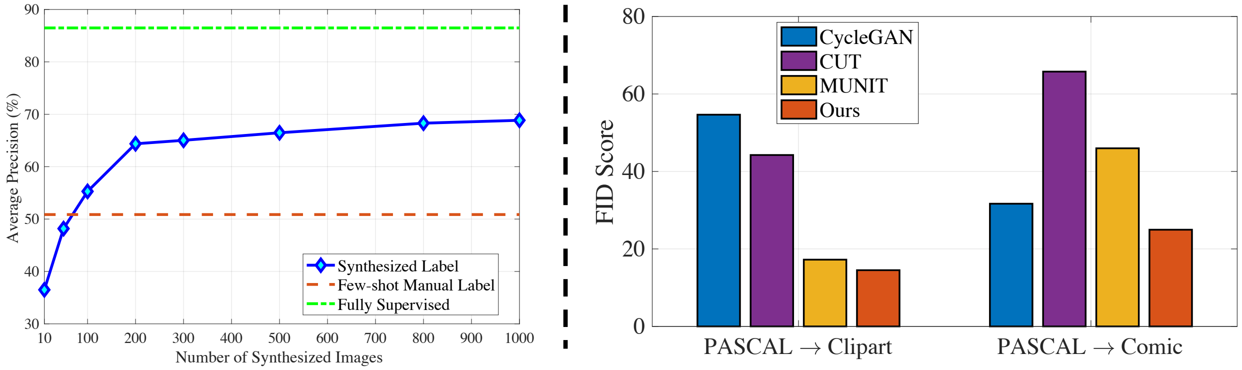
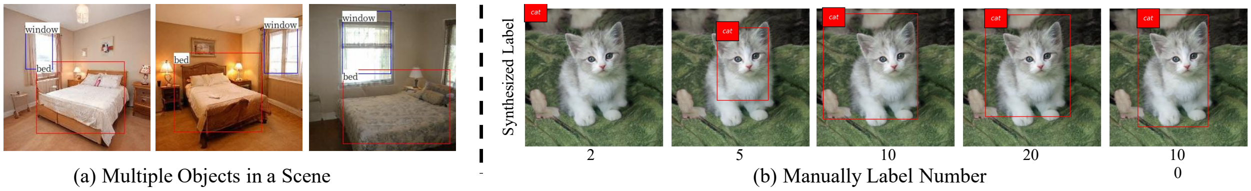


Figure 4: Left: Object detection performance with different numbers of synthesized images and annotations, under the setting of Table 1. Right: Synthesized images quality comparison between our proposed data factory and other image translation based methods, measured with FID score (↓).



(a) Multiple Objects in a Scene

(b) Manually Label Number

Figure 5: (a) Multiple objects "window" and "bed" are synthesized and labeled in a single indoor scene image. (b) Manual label number for training label synthesis branch study. It is observed that increasing manual label number for training label synthesis branch helps generate more precise bounding box and detect more accurate objects.

Multiple Objects in a Single Image. In order to further prove the effectiveness of our proposed method for the scenario that there are multiple objects in a single image, we conduct the experiment where the indoor scene images are synthesized and the corresponding "window" and "bed" objects label are generated. The qualitative experimental results are shown in Fig. 5. It is shown that our method can effectively synthesize and label the image where multiple objects appear in a single scene.

Manual Label Number Study. In order to study the influence of the manual label number for training label synthesis branch, we change the manual label number and train the efficient labeled data factory on the cat category, whose label synthesized results are shown in Fig. 5. It is proven that increasing manual label number for training label synthesis branch helps generate more precise bounding box and detect more accurate objects.

5 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. The proposed approach is validated in various settings and surpasses other methods, demonstrating its effectiveness for SF-FSDA problem.

References

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

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

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

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

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

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.

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

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.

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

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.

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.

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

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

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

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

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.

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.

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*, 2020a.

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*, 2020b.

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

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

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

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

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

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

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

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

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.

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

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

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.

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

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*, 2020b.

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

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

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.

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.

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

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.

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

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.

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.

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

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

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

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.

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

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

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

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.

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.

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

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

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

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.

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

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.

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.

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

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.

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

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

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.

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.

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.

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

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

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.

Appendix

In this supplementary material, we provide the additional information for,

- **A** detailed framework implementation and training parameters,
- **B** detailed information about datasets involved in the experiments,
- **C** additional qualitative experimental results,
- **D** additional quantitative experimental results.

A Framework Implementation

This section provides detailed implementation of the image synthesis and label synthesis branches of our proposed efficient labeled data factory, and their corresponding training details.

A.1 Image Synthesis Branch

Network Structure. The network is built upon the structure of StyleGAN2 Karras et al. (2020b), with additional dual-generator and dual-discriminator design to incorporate extra text guidance and relaxed realism strategy for avoiding overfitting to few-shot image samples from the target domain.

To bring in the text guidance Gal et al. (2021), the dual-generator strategy is deployed. We have two generators, G_t to be adapted, and a frozen G for image changing reference. The second generator G follows the same structure as the original one, and remains frozen during the whole training process to keep generating original images without adaptation.

In order to prevent the synthesized images from overfitting to the few-shot image samples of the target domain, the dual-discriminator strategy is deployed, introducing the relaxed realism Ojha et al. (2021). As introduced in Sec. 3.2 of the main paper, the idea is to only discriminate the fake and real images on the image level with the discriminator D_{img} when the corresponding latent code z is sampled from a small anchor region. The image-level discriminator D_{img} follows the original design as in Karras et al. (2020b). When z is sampled outside of the region, we compute the discriminator loss only on the patch level with the second discriminator D_{patch}, which is the subset of the discriminator D_{img}. The two discriminators are trained 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 λ_f, which indicates the frequency of sampling from the anchor region and computing the image-level loss instead of the patch-level loss.

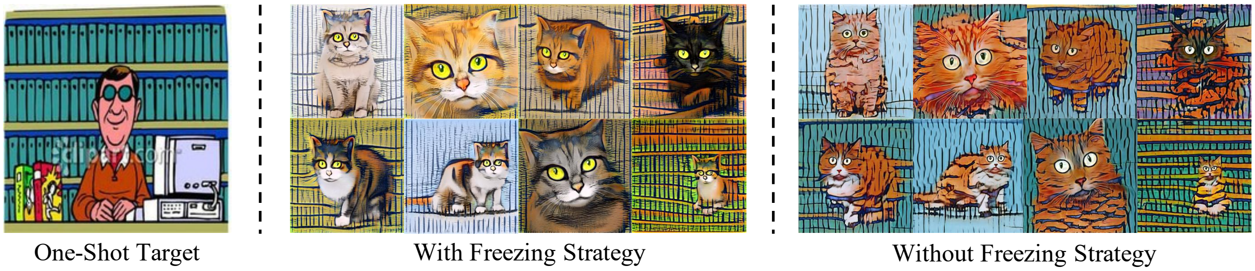
Training Parameters. We adjust the weight of image guidance λ_1 and the weight of text guidance λ_2 in Eq. (5) of the main paper under different scenarios, to balance different guidances. Under the PASCAL VOC→Clipart setting, we set both weights to 1.0. Under the PASCAL VOC→Comic setting, we increase the weight of text guidance to 5.0. Another parameter is the sampling frequency from the randomly sampled small anchor region, λ_f, which decides how often we compute the discriminator loss on the whole image level. We set the frequency to 2, alternatively computing the loss on the image level and the patch level. The rest training details follow the StyleGAN2 Karras et al. (2020b) with the augmentation strategy introduced in Karras et al. (2020a). The training iteration for image synthesis is set as 1000.

In order to avoid overfitting to the few-shot image samples from the target domain, we develop the freezing strategy for image synthesis training (see the *image synthesis training strategy* part in Sec. 3.2 of the main paper). More specifically, we adapt the generator G_t only on specific 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 discriminator, D_{img} and D_{patch}, except for the final layer.

A latent code $\mathbf{z}$ is randomly sampled for each image with a dimension of 512 for detection labeling. In order not to generate low quality images, the latent code is truncated by the average latent code to avoid sampling from low probability density region. More details can be found in Sec. B "Truncation trick in W" of Karras et al. (2019).

A.2 Label Synthesis Branch

Network Structure. Our label synthesis branch is built by utilizing StyleGAN2 Karras et al. (2020b) generator acquired in the image synthesis step as the backbone network, and then adding different prediction heads on this basis.


One-Shot Target

With Freezing Strategy

Without Freezing Strategy

Figure 6: Qualitative results comparison, with/without freezing strategy for image synthesis training, under the one-shot target domain condition. It is shown that the freezing strategy can help to improve the image generation diversity and to prevent overfitting to the one-shot target domain effectively.

For the backbone network, we take the intermediate feature map 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 three prediction heads for keypoint, offset, and bounding box size predictions, respectively. The prediction head is composed of, 3×3 convolutional layer, ReLU, and 1×1 convolutional layer.

Training Parameters. We mainly follow the training parameters and details in Zhou et al. (2019). We set the hyper-parameters in Eq. (7) to $\lambda_{off} = 1.0$ and $\lambda_{size} = 0.5$. 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.

B Datasets Information

In Sec. 4.1 of the main paper, we provide the information about the datasets which are involved in our experiments. We here further provide additional datasets information.

PASCAL VOC. PASCAL VOC 2007 & 2012 dataset Everingham et al. (2010) contains natural images. 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. 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.

Clipart1k. Clipart1k dataset includes the clipart images collected from the CMPlaces dataset Castrejon et al. (2016) and two image search engines Inoue et al. (2018). Clipart1k is meant for the education and research purposes only.

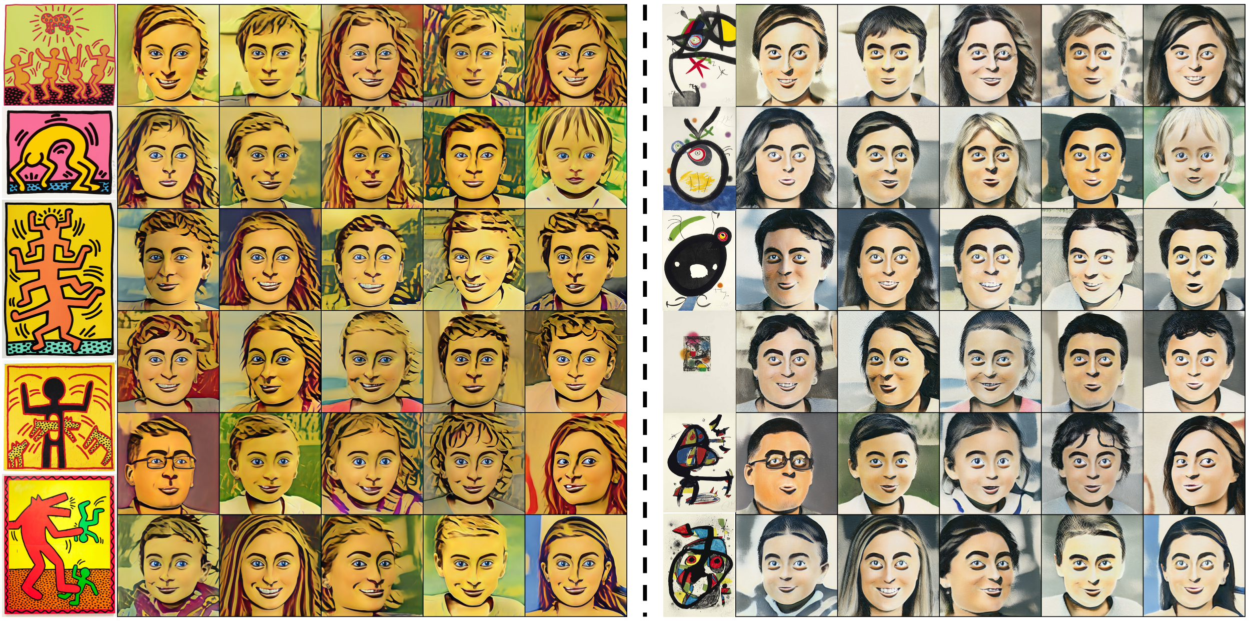
Comic2k. Comic2k dataset covers the comic images collected from BAM! Wilber et al. (2017). Comic2k dataset is meant for the education and research purposes only.

Watercolor2k. Watercolor2k dataset contains the watercolor images collected from BAM! Wilber et al. (2017), which are meant for the education and research purposes only.

C Additional Qualitative Results

C.1 Freezing Strategy for One-Shot Target Domain

In Fig. 5 (k)-(l) of the main paper, we show the qualitative comparison results for the ablation study with/without the freezing strategy of image synthesis training. In order to further prove the validity of the freezing strategy under the extreme case, we here provide the qualitative comparison in Fig. 6 under the one-shot target domain condition, *i.e.*, there is only one image available on the target domain. From Fig. 6, 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.



Clipart (Guided by Few-shot Keith Haring Style Images)

Clipart (Guided by Few-Shot Joan Miro Style Images)

Figure 7: Image synthesis results on human face. The image synthesis is guided by the text guidance “clipart” and the few-shot image samples guidance from “Keith Harring” and “Joan Miro” style paintings. The first column of the left part and the right part is the example of the “Keith Harring” and “Joan Miro” painting style.

C.2 Additional Image Synthesis Results on Human Face

In Sec. 4 of the main paper, we provide the experimental results for SF-FSDA under the PASCAL VOC→Clipart, Comic settings, to synthesize the cat and car objects. In order to further prove the ability of our data factory for image synthesis, we utilize our proposed data factory to synthesize other objects, human face, guided by the few-shot images and the text. The few-shot image guidance includes different artistic style paintings and the provided text guidance is “clipart”. The qualitative results shown in Fig. 7 prove that our proposed data factory effectively synthesizes the target-domain like images under the text and the few-shot image samples guidance.

C.3 Additional Image Synthesis Results on Adverse Weather Simulation

In order to show the possible application of our method to different scenarios, *e.g.*, autonomous driving, we conduct experiments adapting original images of the car category to reflect adverse weathers, *e.g.*, snowy weather. Fig. 3 shows the adapted synthesized images under adverse weather conditions. For the adaptation, text description from the “sunny” to the “snowy” and “foggy” together with 5 style images of street scenes under “snowy” and “foggy” weather are provided as guidance. As shown in the Fig. 8, our data factory is effective for synthesizing the adverse weather images.

C.4 Qualitative Results on the Target Domain

In Sec. 4 of the main paper, we show the quantitative robust object detection results on the Clipart1k and Comic2k datasets. In order to further prove the effectiveness of our proposed data factory for robust object detection, we here show the qualitative object detection results on the target domain, *i.e.*, Clipart1k and Comic2k. From the qualitative results in Fig. 9, it is shown that our proposed efficient labeled data factory adapts the source pretrained object detection model to the target domain, and improves the object detection performance on the target domain effectively.

C.5 Additional Baseline Comparison Results

In Fig. 5(d)-(g) of the main paper, we show the qualitative comparison results between our proposed efficient labeled data factory method and other image translation based methods, CycleGAN Zhu et al. (2017), CUT Park et al. (2020),

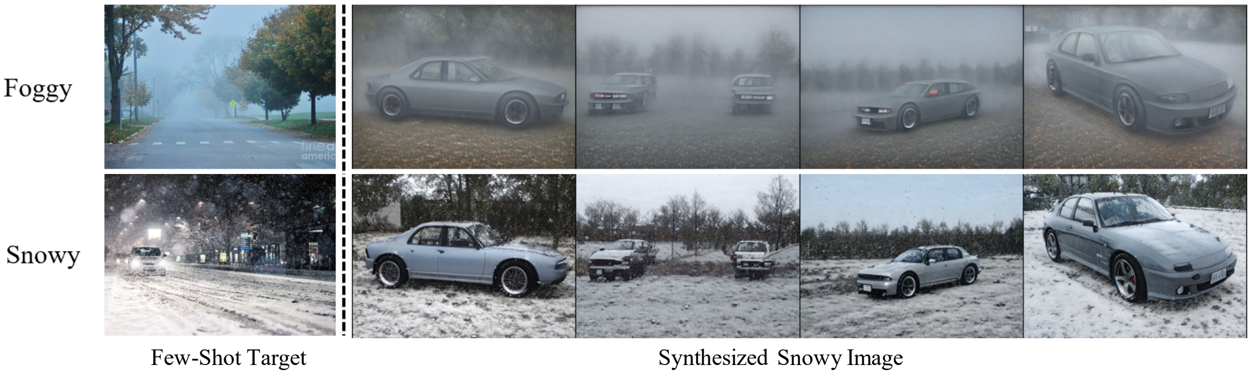


Figure 8: Image synthesis results on car, with few-shot image and text guidance of “snowy” and “foggy”.

and MUNIT Huang et al. (2018). We here provide additional comparison results in Fig. 10 to validate that our proposed efficient labeled data factory can synthesize the images and corresponding object bounding box and category annotations effectively.

C.6 SF-FSDA: PASCAL VOC→Watercolor

In Sec. 4 of the main paper, we show quantitative and qualitative experimental results for SF-FSDA under the PASCAL VOC→Clipart and Comic settings. In order to further verify the validity of our proposed efficient labeled data factory method for SF-FSDA, 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 3 and Table 4 of the main paper, 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 cat and car categories, respectively. In Fig. 11, we show the image and label synthesis results from our data factory, under the PASCAL VOC→Watercolor setting.

C.7 Comparison to StyleGAN-NADA

In order to further validate the effectiveness of our proposed method, we compare our method with recent language guided image synthesis method, StyleGAN-NADA Gal et al. (2021). As shown in Fig. 12, it is observed that the images synthesized by our method are more like Clipart image than StyleGAN-NADA, benefiting from the additional few-shot samples guidance and the freezing training strategy for image synthesis.

C.8 target domain image guidance samples number study

To explore the effect of target samples number for the image synthesis branch, we compare the image synthesis results guided by varying numbers of target samples for training. Our experimental findings, as illustrated in Fig. 13, indicate that our method is able to avoid collapse even with one-shot guidance. However, we observe that using too few examples results in a lack of intra-style differences, ultimately resulting in monochromatic images with a uniform background.

C.9 freezing strategy study

We investigate the freezing strategy hyperparameter choice by experimenting with different sets of frozen parameters. We conduct experiments by unfreezing different numbers of layers of the generator and report our findings in Fig. 14. Specifically, we observe that unfreezing too many layers of the image generator leads to corrupted and blurred generation results, while freezing too many layers limits the model’s flexibility for adaptation, resulting in the collapse to monochromatic local patterns. We also experiment with different freezing strategies for the discriminator and find that unfreezing more layers often leads to imbalanced training. Therefore, for all our experiments, we only unfreeze the last linear layer of the discriminator.

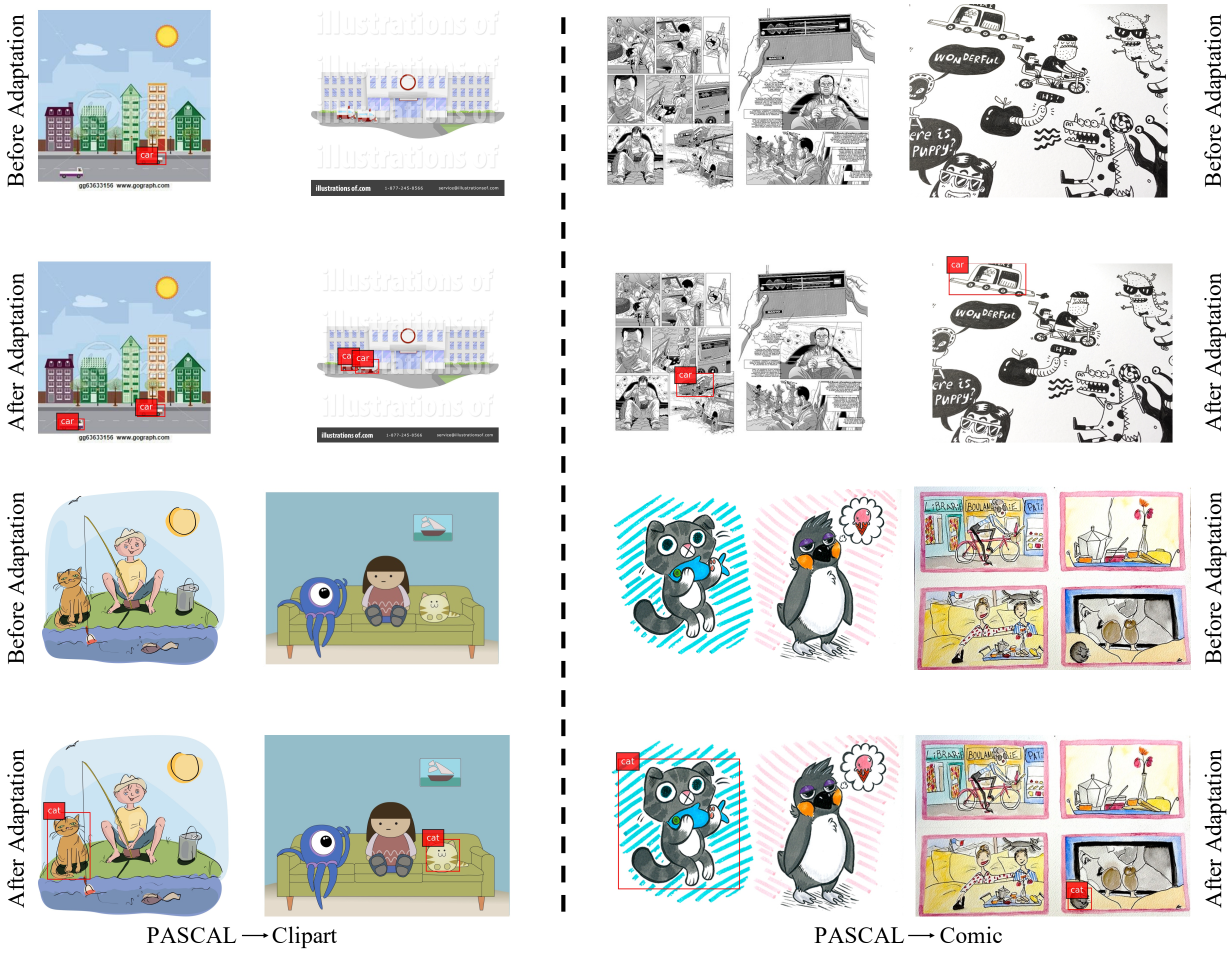


Figure 9: Qualitative object detection results on the target domain, Clipart and Comic. "Before Adaptation" represents the object detection results when applying the source-pretrained object detection model to the target domain. "After Adaptation" shows the object detection results after fine-tuning the source-pretrained model on the synthesized images and labels from our proposed data factory. The image without detected bounding box indicates that the model cannot detect the objects in the image.

D Additional Quantitative Results

D.1 Comparison to Few-Shot Labeled Target Domain Images

In order to further prove the effectiveness of our proposed approach, we compare our approach to the few-shot labeled target domain images, where the few-shot target domain images are labeled and then are used to fine-tune the source-pretrained object detection model. The number of manually labeled few-shot target domain images is the same as that of manually labeled synthesized images used for training label synthesis branch in the efficient labeled data factory. Under the PASCAL VOC → Clipart "cat" category setting in Table 3 of the main paper, the few-shot labeled target domain images fine-tuning baseline reaches the performance 29.97%, while our approach reaches 32.50%. It is because our efficient labeled data factory can synthesize a large number of images and corresponding label, while the few-shot labeled target domain samples are insufficient.

D.2 Target Domain Samples Number Study

As observed in Fig. S1 of the supp. and Fig. 3(d) of the main paper, more target samples can improve the diversity of the synthesized image samples and prevent overfitting. Quantitatively, corresponding to Fig. 3(d), the object detection

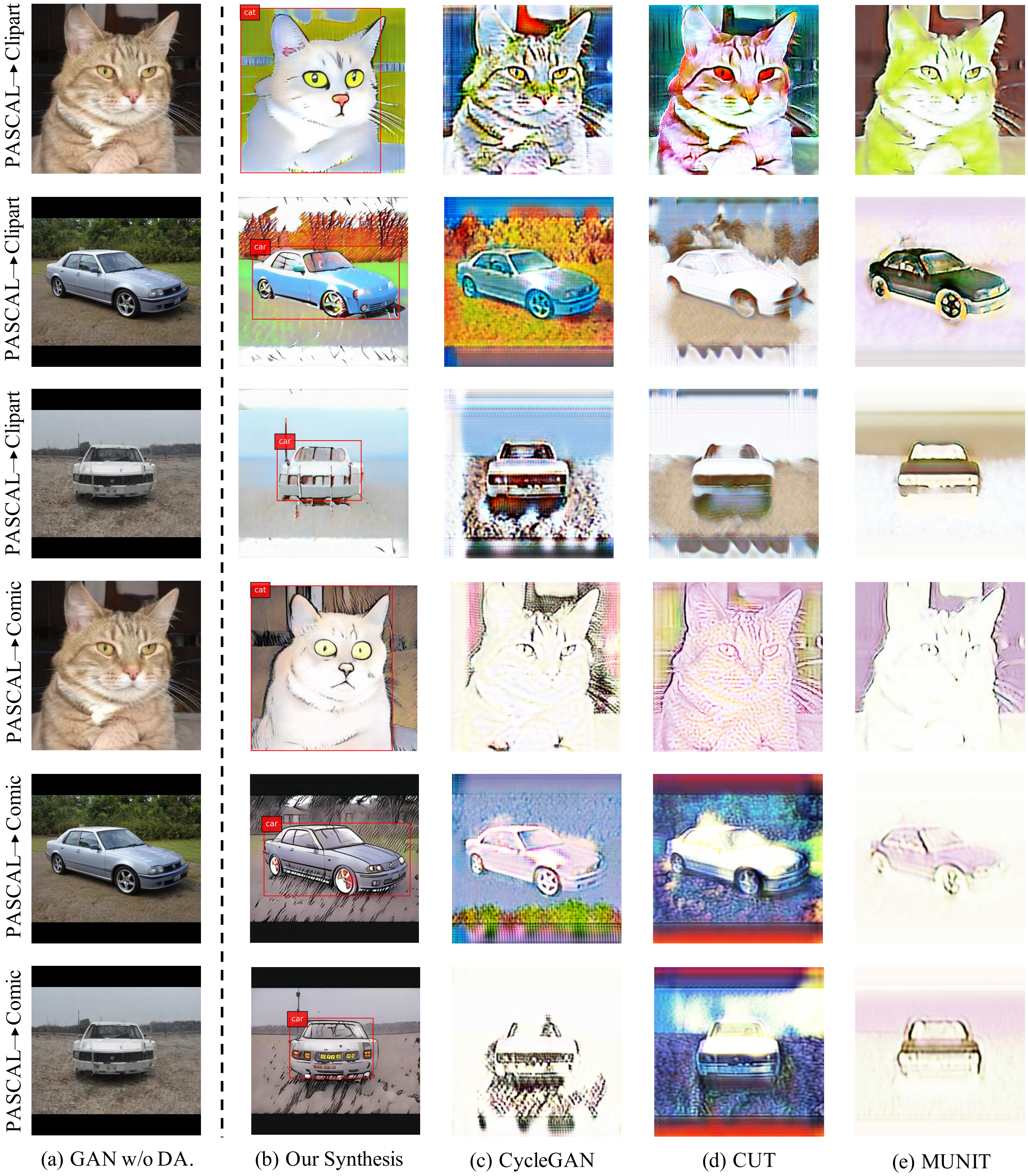


Figure 10: Comparison between data factory and other image translation based methods. (b) is synthesized from our proposed efficient labeled data factory, while (c)-(e) are generated from the image-translation based methods, CycleGAN, CUT and MUNIT.

performance under PASCAL VOC → Clipart "cat" category is 32.50%, outperforming the performance corresponding to Fig. S1, 23.15%.

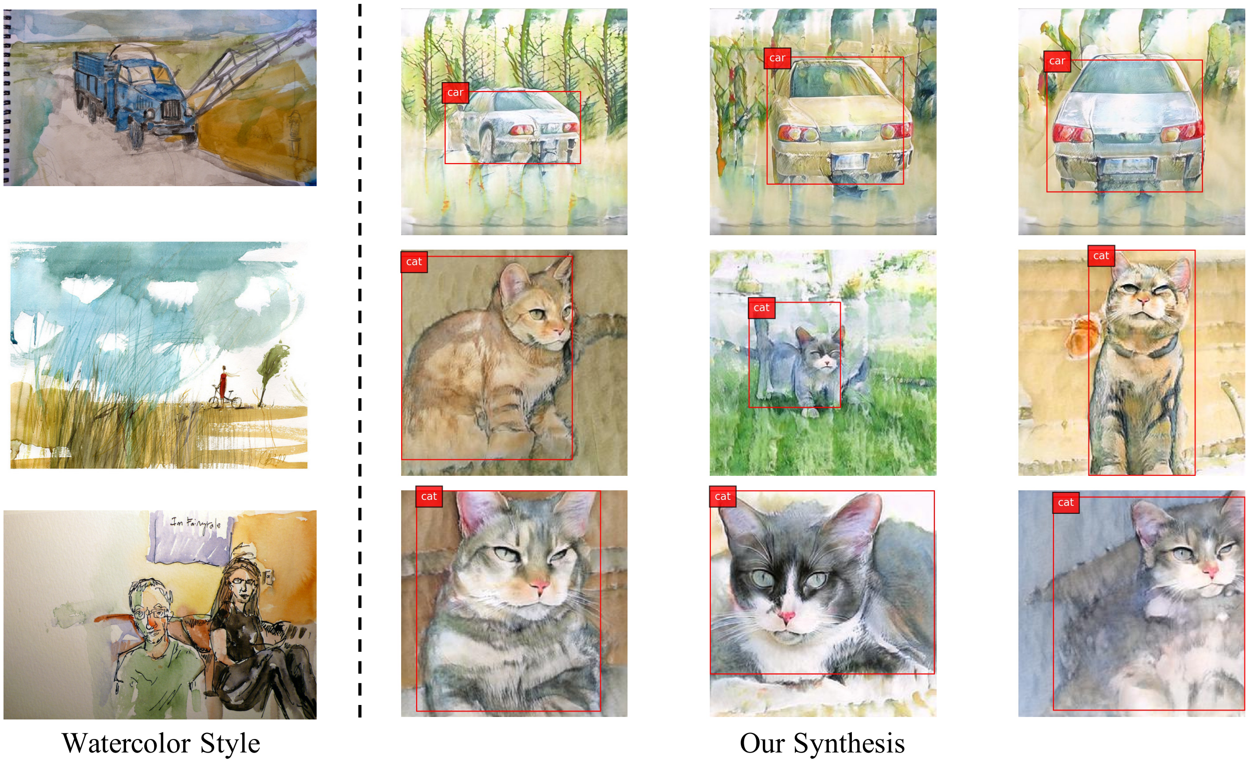


Figure 11: Image and label synthesis results for SF-FSDA, under the PASCAL VOC→Watercolor setting. The text guidance is “watercolor”, and the few-shot image samples guidance is the watercolor image from Inoue et al. (2018).

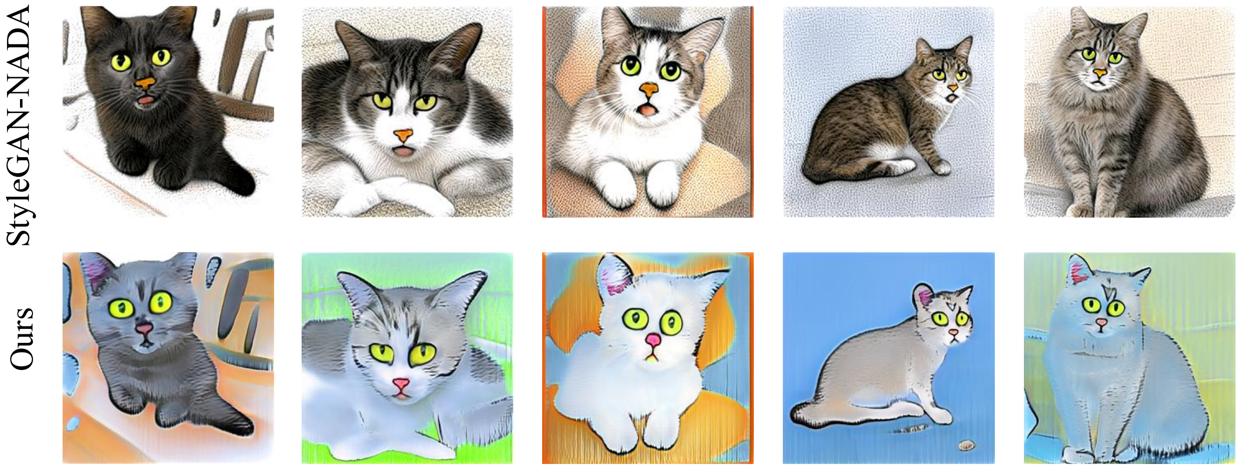


Figure 12: Comparison with StyleGAN-NADA Gal et al. (2021) for Clipart image synthesis.

D.3 Multi-domain study

To evaluate the effectiveness of our method in multi-target domain scenarios, we conduct experiments on multiple target domains and report the results in Table 5. Our efficient labeled data factory achieves reliable performance on this task. Specifically, as depicted in the table, we observe that transferring from one target domain to another can even improve the performance compared to direct adaptation from the original source pretrained model. This highlights the potential of our method to be used in real-world scenarios with multiple target domains.

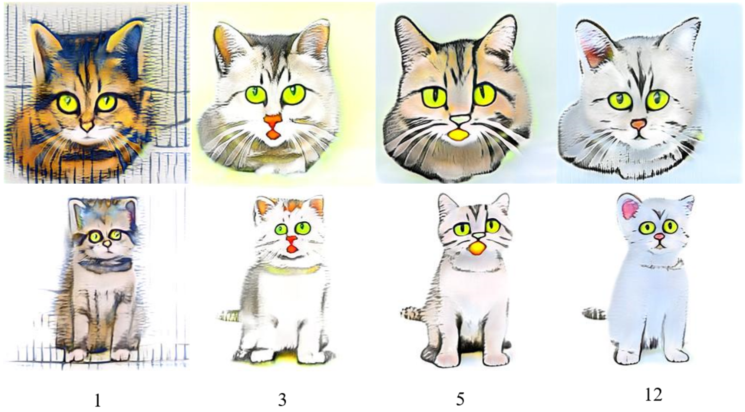


Figure 13: Image synthesis results guided by different number of image guidance samples (1, 3, 5, 12) from the target domain (Clipart).

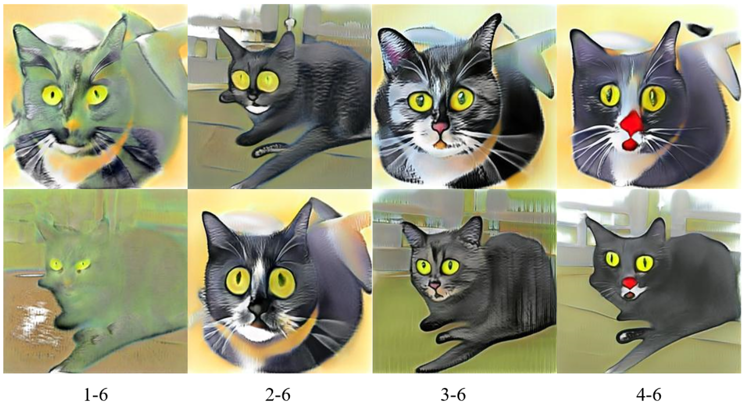


Figure 14: Image synthesis results obtained by fine-tuning different layers of the image generator. Specifically, we denote unfreezing and fine-tuning the first to the last layer of the image generator as 1-6, respectively.

Table 5: Multi-target domain SF-FSDA results. We evaluate our SF-FSDA model's performance on two adaptation scenarios: PASCAL VOC → Comic and PASCAL VOC → Clipart → Comic. In the first scenario, we adapt the PASCAL VOC trained model on Comic-style training samples synthesized by our data factory. In the second scenario, we first adapt the PASCAL VOC trained model to Clipart-style training samples synthesized by our data factory, and then further adapt it to the Comic-style training samples synthesized by our data factory.

Class	PASCAL VOC → Comic	PASCAL VOC → Clipart → Comic
Cat	37.74	**40.14**
Car	54.68	**57.02**

D.4 Manual label samples number study

We investigated the impact of the manual labels number on the SF-FSDA performance, whose results are presented in Table 6. The experimental results reveal that the SF-FSDA performance improves with an increase in the number of manual labels. However, the rate of improvement decreases as the number of manual labels becomes larger. After balancing stability and efficiency, we selected the number of manual labels as 10.

Table 6: SF-FSDA performance on the Comic target domain for Cat category, with varying numbers of manually labeled samples utilized for label synthesis.

Number of Manual Labels				
5	10	20	50	100
37.04	37.74	38.54	39.25	39.87