

# One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models

Yasser Benigim<sup>1,2</sup> Subhankar Roy<sup>1</sup> Slim Essid<sup>1</sup> Vicky Kalogeiton<sup>2</sup> Stéphane Lathuilière<sup>1</sup>

<sup>1</sup> LTCI, Télécom-Paris, Institut Polytechnique de Paris

<sup>2</sup> LIX, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris

yasser.benigim@telecom-paris.fr

## Abstract

Adapting a segmentation model from a labeled source domain to a target domain, where a single unlabeled datum is available, is one of the most challenging problems in domain adaptation and is otherwise known as one-shot unsupervised domain adaptation (OSUDA). Most of the prior works have addressed the problem by relying on style transfer techniques, where the source images are stylized to have the appearance of the target domain. Departing from the common notion of transferring only the target “texture” information, we leverage text-to-image diffusion models (e.g., Stable Diffusion) to generate a synthetic target dataset with photo-realistic images that not only faithfully depict the style of the target domain, but are also characterized by novel scenes in diverse contexts. The text interface in our method *Data AugmenTation with diffUision Models (DATUM)* endows us with the possibility of guiding the generation of images towards desired semantic concepts while respecting the original spatial context of a single training image, which is not possible in existing OSUDA methods. Extensive experiments on standard benchmarks show that our DATUM surpasses the state-of-the-art OSUDA methods by up to +7.1%. The implementation is available at : <https://github.com/yasserben/DATUM>

## 1. Introduction

Semantic segmentation (SS) is one of the core tasks in computer vision [8, 53, 56], where a neural network is tasked with predicting a semantic label for each pixel in a given image [11]. Given its importance, SS has received significant attention from the deep learning community and has found numerous applications, such as autonomous driving [6, 7], robot navigation [26], industrial defect monitoring [34].

The task of semantic segmentation is known to require pixel-level annotations which can be costly and impractical in many real-world scenarios, making it challenging to train segmentation models effectively. Moreover, the issue of *domain shift* [46] can cause segmentation models to underper-

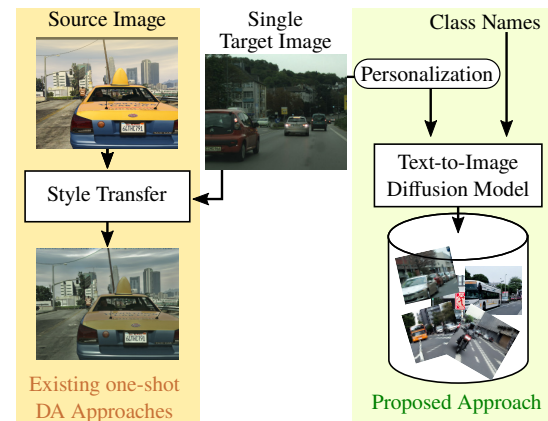


Figure 1. In existing OSUDA methods data augmentation is done via stylization [16, 29]. In our proposed approach, we prompt the text-to-image diffusion models [36] to generate new images that not only depict the style of the target domain, but also more faithfully capture the diversity of the scene content.

form during inference on unseen domains, as the distribution of the training data may differ from that of the test data.

To make learning effective without needing annotations on the target domain, several Unsupervised Domain Adaptation (UDA) methods have been proposed for the task of semantic segmentation [10, 20, 47, 52]. Fundamentally, the UDA methods collectively use the labeled (or *source*) and the unlabeled (or *target*) dataset to learn a model that works well on the target domain. Despite being impressive in mitigating the domain gap, the UDA methods rely on the assumption that a considerably large dataset of unlabelled images is at disposal. However, collecting a large target dataset before adaptation poses as a bottleneck in the rapid adoption of segmentation models in real-world applications. To circumvent this issue, several works have investigated the feasibility of using just a small subset of the unlabeled target samples (at times just *one* sample) to adapt the model. This adaptation scenario is known as One-Shot Unsupervised Domain Adaptation (OSUDA) [3, 16, 29, 55], where, in addition to the source dataset, only a single *unlabelled* target sample is available.

While the OSUDA setting is realistic and cost-effective,

relying solely on a single target image makes it challenging for traditional UDA methods to estimate and align distributions. To address the lack of target data, the OSUDA approaches generally overpopulate the target dataset with source images *stylized* as *target-like* ones [16, 29]. Albeit effective, these methods result in a target dataset that is limited to the scene layouts and structures inherent to the source dataset (Fig. 1 left). In this work, we argue that simply mimicking the style of the target is insufficient to train a robust target model, especially when only limited information about the target domain is available. Thus, we seek for diversifying the scene content and spatial layout, more than what the source images can offer. Moreover, generating high-fidelity images is yet another challenging problem. Thus, in this work, we focus on denoising diffusion models (DM) [19, 36], a family of generative models with excellent capability in generating high-quality images. We propose to leverage DMs to augment the target dataset with images that not only resemble the target domain at hand, but also contain diverse and plausible scene layouts due to rich prior knowledge encoded in DMs (see Fig. 1 right).

In detail, we fine-tune a DM [19, 36] on the single target sample to generate an auxiliary large target dataset. Following recent work [14, 40], we represent the target image with a special, rare and unique *token* that encapsulates its visual appearance. Then, we exploit the vast knowledge of DMs about the objects (or *things* classes) present in the source domain for a driving scenario [13, 28, 41]. Specifically, we prompt the model to generate a target dataset depicting such objects in a multitude of scenes, while maintaining the appearance tethered to the overall target domain *style* via the unique token. Once an augmented target dataset is made available, any UDA method can be used to adapt to the target domain. We thus present our method **Data AugmenTation with diffUision Models (DATUM)**, for addressing OSUDA, as a connotation to the setting of having access to a single “datum” from the target domain. Our approach has the advantage of making any UDA method compatible with the one/few-shot setting. In our experiments, we add DATUM to existing UDA methods and compare against the state-of-the-art OSUDA. Our results and analysis demonstrate the efficacy of DATUM and its ability to diversify the target dataset. We believe that DATUM can contribute significantly to semantic segmentation as a *plug-and-play* module.

Our **contributions** are three-fold: (i) We demonstrate, for the first time in the context of SS, the importance of generating semantically diverse and realistic target-like images in OSUDA. (ii) We propose DATUM, a generic data augmentation pipeline powered by DMs, for addressing the challenging yet relevant task of OSUDA, and (iii) while being conceptually simple, we show with extensive experiments, on standard sim-to-real UDA benchmarks, that DATUM can easily surpass the state-of-the-art OSUDA methods.

## 2. Related Works

**Unsupervised domain adaptation.** To bridge the domain gap between the source and target datasets, unsupervised domain adaptation (UDA) methods have been proposed, which can be roughly categorized into three broad sub-categories depending on the level where the distribution *alignment* is carried out in the network pipeline. First, the feature-level alignment methods aim at reducing the discrepancy between the source and target domains in the latent feature space of the network under some metric. As an example, these methods include minimizing the Maximum Mean Discrepancy (MMD) [4] or increasing the domain confusion between the two domains with a discriminator network [21, 24, 30, 43, 52]. The latent space being high dimensional, the second category of UDA methods [32, 48, 49, 52] exploits the lower dimensional and more structured output space of the network to mitigate domain shift, while borrowing *e.g.*, adversarial alignment techniques. The third category includes methods [20, 21, 24, 27, 30, 43, 45, 52] that align the source and the target domains in the input (or pixel) space by generating target-like source images via style transfer [15, 25, 61]. There is yet another successful line of UDA works that exploit self-training using a student-teacher framework [2, 22, 23].

While the above UDA methods are effective under the standard adaptation setting to varying degrees, where the entire target dataset is available for training, style transfer-based methods are particularly effective when the target data is inadequate to approximate a distribution. Different from the existing methods [3, 16], which are just capable of transferring style (or “appearance”) information to the source images, our proposed DATUM can additionally generate novel and structurally coherent content in the target domain.

**Few-shot adaptation.** To improve the sample efficiency of the (UDA) methods, *supervised* few-shot domain adaptation (FSDA) methods [12, 31, 54] relax the need of having a large unlabeled target dataset, in favour of assuming access to a few but *labeled* samples of the target domain. The FSDA methods [57, 60] exploit the labeled target samples to construct prototypes to align the domains. The setting of OSUDA is a more challenging version of FSDA, where a *single* target sample is available *without* any annotation. Due to the lack of means of constructing prototypes or aligning distributions with a single target sample, OSUDA methods [16, 29, 55] are based on transferring style from the target sample to the source dataset to artificially augment the target dataset. Once augmented, UDA methods such as self-training [16], consistency training [29], prototypical matching [55], are applied. Similar to [16], we use the self-training framework DAFormer [22] to adapt to the generated target images. However, unlike the prior OSUDA works [16, 29, 55], DATUM’s data generation pipeline is stronger, conceptually simpler and does not rely on many heuristics.

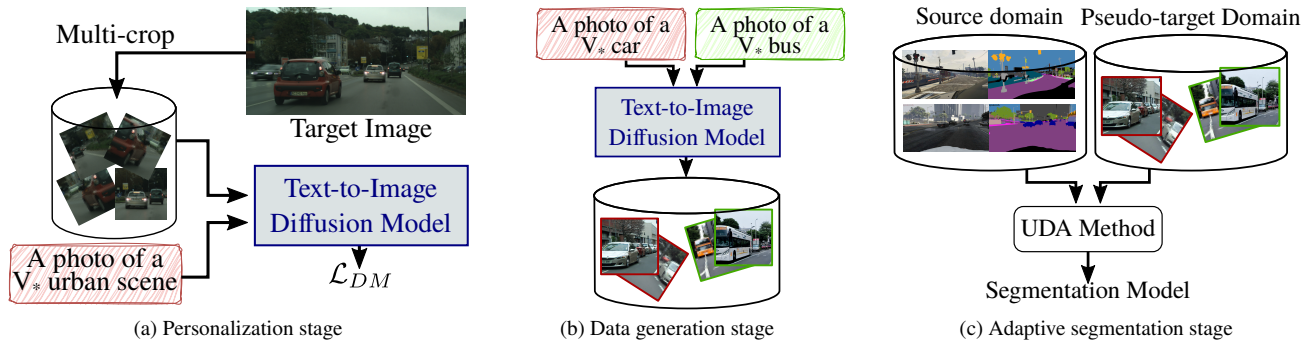


Figure 2. **The three stages of DATUM.** In the **personalization stage** (a), we learn to map a unique token  $V_*$  with the appearance of the target domain using a single target image. In the **data generation stage** (b), we employ the personalized model to generate a large dataset corresponding to the target distribution. Class names are used to improve diversity. Finally, the **adaptive segmentation stage** (c) consists in training an existing UDA framework on the labeled source and the generated unlabeled pseudo-target datasets

**Diffusion models.** Very recently, diffusion models (DM) [19, 44] have brought a paradigm shift in the generative modeling landscape, showing excellent capabilities at generating photo-realistic text-conditioned images [33, 36, 41]. To allow personalized and more fine-grained generation, works such as DreamBooth [40], Textual Inversion [14] and ControlNet [58] have extended DMs with different levels of fine-tuning, offering more flexibility. However, a handful of recent works [1, 18, 42] has explored the possibility of using a latent diffusion model [36], a fast alternative to DM, for generating class-conditioned *synthetic* datasets, as replacements of the *real* counterparts, to solve image recognition tasks. In contrast to these approaches, we specifically address the problem of domain adaptation by augmenting the target domain. We adopt a fine-tuning strategy [40] that explicitly incorporates the appearance of the target domain. Our approach associates a word identifier with the content of the target image, resulting in high-fidelity synthetic generations.

### 3. Method

In this work, we propose **Data AugmenTation with diffUision Models (DATUM)**, a generic method for creating *synthetic* target dataset by using a single real sample (and hence, *one-shot*) from the target domain. The synthetic dataset is then used for adapting a segmentation model. Sec. 3.1 introduces the task and gives a background about DM, while Sec. 3.2 describes DATUM.

#### 3.1. Preliminaries

**Problem formulation.** In this work, we address the problem of One-Shot Unsupervised Domain Adaptation (OSUDA), where we assume access to  $N^S$  labeled images from a source domain  $D^S = \{(X_i^S, Y_i^S)\}_{i=1}^{N^S}$ , where  $X_i^S \in \mathbb{R}^{H \times W \times 3}$  represents an RGB source image and  $Y_i^S \in \mathbb{R}^{H \times W \times |\mathcal{C}|}$  the corresponding one-hot encoded ground-truth label, with each pixel belonging to a set of  $\mathcal{C}$  classes. Unlike, traditional UDA methods [20, 52], in OSUDA we have access to a *single*

*unlabeled* target sample  $X^T$ , where  $X^T \in \mathbb{R}^{H \times W \times 3}$ .

In the context of semantic segmentation, the goal in OS-UDA is to train a segmentation model  $f: \mathcal{X} \rightarrow \mathcal{Y}$  that can effectively perform semantic segmentation on images from the target domain. Given the sheer difficulty in training  $f(\cdot)$  with the single target image, our method seeks to generate a synthetic target dataset by leveraging a text-to-image DM.

**Background on Diffusion Models.** Diffusion Models (DM) [19] approach image generation as an image-denoising task. We obtain a sequence of  $T$  noisy images  $X_1, \dots, X_T$  by gradually adding random Gaussian noises  $\epsilon_1, \dots, \epsilon_T$  to an original training image  $X_0$ . A parameterized neural network  $\epsilon_\theta(\cdot, t)$  is trained to predict the noise  $\epsilon_t$  from  $X_t$  for every denoising step  $t \in \{1, \dots, T\}$ . Denoising is typically carried out with a U-Net [38]. To enable conditioning, the network  $\epsilon_\theta(X_t, y, t)$  is conditioned on an additional input  $y$ . In the case of text conditioning, the embeddings from a text-encoder  $\tau_\theta$  for the text  $y$  are used to augment the U-Net backbone with the cross-attention mechanism [50]. For a given image-caption pair, the conditional DM is learned using the following objective:

$$\mathcal{L}_{DM} = \mathbb{E}_{X, y, \epsilon \sim \mathcal{N}(0, 1), t} \left[ \|\epsilon - \epsilon_\theta(X_t, t, \tau_\theta(y))\|_2^2 \right] \quad (1)$$

To improve efficiency, we employ a DM, which operates in the latent space of a pre-trained autoencoder [36].

#### 3.2. Data Augmentation with Diffusion Models

Our proposed DATUM works in three stages and is shown in Fig. 2. In the first stage, called the **personalization stage**, we fine-tune a pre-trained text-to-image DM model by using multiple crops from the single target image (see Fig. 2a). This steers the DM towards the distribution of the target domain of interest. Next, in the second **data generation stage**, we prompt the just fine-tuned text-to-image DM to generate a synthetic dataset that not only appears to be sampled from the target domain, but also depicts desired semantic concepts (see Fig. 2b). Finally, the **adaptive segmentation stage**

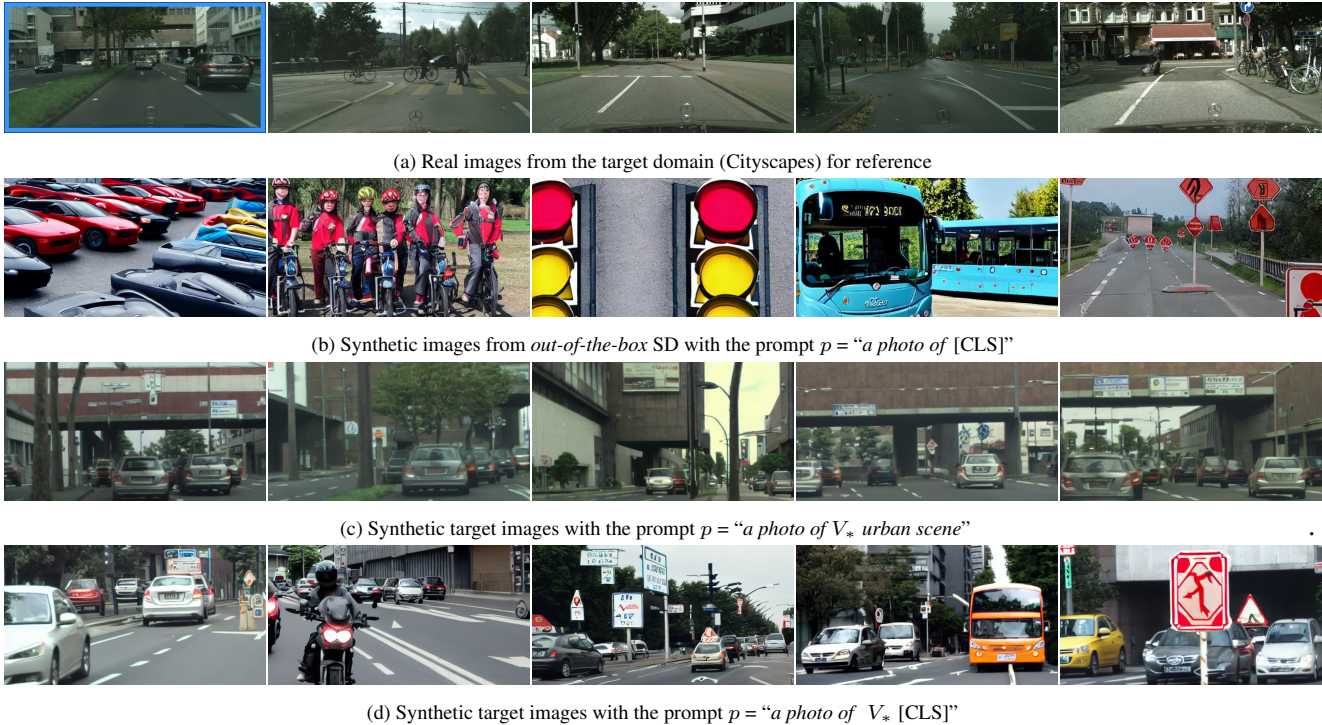


Figure 3. **Qualitative study illustrating the underlying motivations of our three-stage approach.** (a) Real images from the Cityscapes target domain. (b) Out-of-the-box Stable Diffusion (SD) can generate photo-realistic images given the [CLS] name in the prompt, but barely have any resemblance to Cityscapes. (c) Fine-tuning SD on a single target image (personalization stage) leads to generations that truly mimic the Cityscapes domain, but at the cost of losing diversity. (d) Our proposed prompting strategy (data generation stage) leads to synthetic generations that are both photo-realistic and also resembles Cityscapes-like images. The blue-framed image in (a) is the training image used to generate the images in rows (c) and (d).

culminates the three stage pipeline of DATUM, where we combine the labeled source data with the synthetic pseudo-target data and train with a general purpose UDA method (see Fig. 2c). Next, we describe each stage in detail.

**Personalization stage.** The goal of the personalization stage is to endow the pre-trained DM with generation capabilities that are relevant to the downstream task. This stage is crucial because simply generating out-of-domain photo-realistic images is not useful for the downstream task. As an example, as shown in Fig. 3(b), when an out-of-the-box DM is prompted with  $p = \text{"a photo of [CLS]"}$ , where CLS represents a user-provided object class from the dataset, the DM generates high-fidelity images that truly depict the desired semantic concept. However, when compared to the real target domain (see Fig. 3(a)) the DM generated images of Fig. 3(b) have little to no resemblance in appearance. Given that the labeled source dataset already provides a rich prior to the segmentation model about the object classes of interest, having more unrelated and unlabeled images is unappealing.

Thus, we strive to imprint the appearance of the target domain into the synthetic dataset, while just using a single real target sample, in order to obtain more targetted synthetic data. Towards that end, we use DreamBooth [40], a recently

proposed technique for fine-tuning the DM, that allows for the creation of novel images while staying faithful to the user-provided subset of images. In detail, DreamBooth associates a unique identifier  $V_*$  to the subset of images as provided by the user by fine-tuning the DM weights. Similarly, we fine-tune the DM on the single target image while conditioning the model with the prompt  $p = \text{"a photo of } V_* \text{ urban scene"}$ . This results in the unique identifier  $V_*$  capturing the target domain appearance. Once trained, we prompt the fine-tuned DM with  $p = \text{"a photo of } V_* \text{ urban scene"}$  and report the results in the Fig. 3(c). We observe the stark improvement in the overall visual similarity with the reference target domain images depicted in Fig. 3(a). As a result of the personalization step with  $V_*$ , we can now condition the DM to generate more samples of the desired target domain.

However, a thorough inspection of the generated images in Fig. 3(c) reveals that the images lack diversity. The DM overfits to the single target image and loses its ability to generate many other objects. For instance, some classes (such as *car*) are repeated whereas others (such as *bus*, *bike*, *truck*) never appear. To prevent this overfitting issue, we train the DM for a limited number of iterations. Moreover, we disable the class-specific *prior-preservation* loss used in Dreambooth [40], designed for not forgetting other concepts, since

our goal is to capture the essence of the target domain, rather than generating a desired object in many unrealistic and unnatural scenarios. For fine-tuning the DM, we optimize the training objective described in Eq. (1).

**Data generation stage.** In the post personalization stage, our goal is to generate a dataset of synthetic images of the target domain. As we use just a single target image in the personalization stage, the generation capability of the fine-tuned DM model can still be limited to few scenes. Therefore, to elicit diverse generations from the fine-tuned DM, at inference we use more targeted prompts than the ones used during training. Specifically, we employ class-wise prompts in the form of: “*a photo of a  $V_*$  [CLS]*”. The [CLS] corresponds to the name of the *things* classes (e.g., *bus*, *person*, etc.) we want to generate, as defined in [5]. Our choice of using only the “things” classes is motivated by the fact that in a driving application, the “things” classes mostly co-occur with “stuff” classes (e.g., *building*, *sky*). Thus, explicitly prompting the model to generate stuff classes is redundant. As shown in Fig. 3(d), injecting the “things” class names into the inference prompt leads to an improved diversity in the generations, while staying close to the target domain in appearance. This helps in combating the long-tailed phenomenon of the semantic segmentation datasets, where some minority classes (e.g., *bike*) appear less frequently than others, such as *cars*, and *road*.

**Adaptive segmentation stage.** While the pseudo-target images in the synthetic dataset contain the user-desired object, they still lack pixel-level information. To overcome this limitation, we resort to UDA techniques that enable a segmentation model to be adapted to an unlabeled target dataset. In this work, we leverage UDA methods such as DAFormer [22] and HRDA [23], but our approach is not exclusive to these two methods. Notably, the optimization objective of these two UDA methods remain unaltered. In summary, our proposed DATUM can transform any UDA method into an effective OSUDA method.

## 4. Experiments

### 4.1. Experimental set up

**Dataset and settings.** We follow the experimental settings established in the OSUDA literature [16, 29, 55] and conduct experiments on two standard *sim-to-real* benchmarks: GTA  $\rightarrow$  Cityscapes and SYNTHIA  $\rightarrow$  Cityscapes, where GTA [35] and SYNTHIA [39] are the source domains in the respective settings, and Cityscapes [9] is the target domain. In details, the GTA dataset comprises 24,966 synthetic images with a resolution of  $1914 \times 1052$ . and SYNTHIA contains 9400 synthetic images of resolution  $1280 \times 760$ . Cityscapes contains 2975 training images and 500 validation images of size  $2048 \times 1024$ , is captured under real-world driving conditions. Note that, since we operate in the one-

shot adaptation scenario, as in [16, 29, 55], we assume to have access to a *datum* from the target domain, which is chosen at random during training.

**Implementation details.** We employ the Stable Diffusion (SD) implementation of Latent Diffusion Models (LDM) [36]. We use the publicly available *Diffusers* library [51] for all the experiments related to generating synthetic data. In particular, for generating synthetic images in the target domain, we start from the Stable Diffusion v1.4 checkpoint [37] and fine-tune it using the DreamBooth [40] method. We refer the reader to LDM [36] for details about the encoder, U-Net, and decoder architectures.

For fine-tuning SD, we randomly crop patches of  $512 \times 512$  from the original  $2048 \times 1024$  resolution, and use a generic prompt  $p = \text{“}a\ photo\ of\ a\ V_*\ urban\ scene\text{”}$ , given that the target domain Cityscapes was captured in an urban set-up [9]. We train SD for 200 iterations, and we find that for the one-shot setting longer training leads to overfitting on the target scene. Once trained, we generate a synthetic target dataset of cardinality 2975, which is equivalent in size to the Cityscapes training set, by utilizing inference prompts of the form  $p = \text{“}a\ photo\ of\ a\ V_*\ [CLS]\text{”}$ . DreamBooth generates images at the same resolution as the input, which is  $512 \times 512$ . This generated dataset then serves as the target domain for adaptation, as in UDA.

For training the final segmentation model on the source and generated datasets, we use the network architecture from state-of-the-art UDA methods [22] that use MiT-B5 [56] as the encoder and a context-aware fusion [22] as the decoder. This is analogous to the most popular ResNet-101 [17] as a backbone, and DeepLabV2 [8] as the decoder. We also experiment with another UDA method: HRDA [23]. For both these experiments with DAFormer and HRDA, we keep the training protocol and hyperparameters unchanged. Both the ResNet-101 and MiT-B5 are pre-trained on ImageNet-1k.

**Evaluation metrics.** Following the standard protocol [29], we report the mean Intersection over Union (mIoU) on the validation set of Cityscapes. For the GTA  $\rightarrow$  Cityscapes benchmark, we compute mIoU over 19-classes, whereas for SYNTHIA  $\rightarrow$  Cityscapes, we report both mIoU<sup>13</sup> and mIoU<sup>16</sup> for 13 and 16 classes, respectively [16].

### 4.2. Comparison with the state-of-the-art

**Baselines.** We compare our proposed method with the state-of-the-art OSUDA methods and UDA methods adapted to the OSUDA setting: CycleGAN [61], ASM [29], OST [3], CACDA [16], SMPPM [55], DACS [47] which are also methods based on data augmentation, as well as ProDA [59], CBST [52], AdaptSeg [48], DAFormer [22] and HRDA [23]. Given that DATUM focuses primarily on data generation, we pair it with the UDA methods DAFormer and HRDA under the OSUDA setting. We denote these models as DAFormer + DATUM and HRDA + DATUM, which use **purely syn-**

Table 1. Comparison with state-of-the-art methods for UDA and OSUDA on the GTA → Cityscapes benchmark. #TS denotes the number of *real* target samples used during training, which are color coded as None, All and One. Methods using ResNet-101 [17] and MiT-B5 [56] are shown in the top and bottom halves, respectively. As an example, DaFormer + DATUM denotes DaFormer trained using the synthetic images generated by our DATUM. \*: results from CACDA [16]; ◊: results from HRDA [23]; and †: results from ASM [29].

	#TS	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
Encoder: ResNet-101																					
Source-Only *	None	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CycleGAN † [61]	All	81.7	27.0	81.7	30.3	12.2	28.2	25.5	27.4	82.2	27.0	77.0	55.9	20.5	82.8	30.8	38.4	0.0	18.8	32.3	41.0
ASM † [29]	All	89.8	38.2	77.8	25.5	28.6	24.9	31.2	24.5	83.1	36.0	82.3	55.7	28.0	84.5	45.9	44.7	5.3	26.4	31.3	45.5
DACS ◊ [47]	All	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
ProDA ◊ [59]	All	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
DAFormer ◊ [22]	All	96.0	72.4	88.0	39.2	37.4	38.0	50.3	54.0	88.4	47.2	89.2	69.8	42.6	88.6	48.6	55.4	0.9	34.6	48.7	57.3
HRDA ◊ [23]	All	96.2	73.1	89.7	43.2	39.9	47.5	60.0	60.0	89.9	47.1	90.2	75.9	49.0	91.8	61.9	59.3	10.2	47.0	65.3	63.0
AdaptSeg * [48]	One	77.7	19.2	75.5	11.7	6.4	16.8	18.2	15.4	77.1	34.0	68.5	55.3	30.9	74.5	23.7	28.3	2.9	14.4	18.9	35.2
CBST * [62]	One	76.1	22.2	73.5	13.8	18.8	19.1	20.7	18.6	79.5	41.3	74.8	57.4	19.9	78.7	21.3	28.5	0.0	28.0	13.2	37.1
CycleGAN * [61]	One	80.3	23.8	76.7	17.3	18.2	18.1	21.3	17.5	81.5	40.1	74.0	56.2	38.3	77.1	30.3	27.6	1.7	30.0	22.2	39.6
OST * [3]	One	84.3	27.6	80.9	24.1	23.4	26.7	23.2	19.4	80.2	42.0	80.7	59.2	20.3	84.1	35.1	39.6	1.0	29.1	23.2	42.3
SMPPM * [55]	One	85.0	23.2	80.4	21.3	24.5	30.0	32.0	26.7	83.2	34.8	74.0	57.3	29.0	77.7	27.3	36.5	5.0	28.2	39.4	42.8
ASM † [29]	One	86.2	35.2	81.4	24.2	25.5	31.5	31.5	21.9	82.9	30.5	80.1	57.3	22.9	85.3	43.7	44.9	0.0	26.5	34.9	44.5
DAFormer [22]	One	85.5	31.2	81.7	24.0	25.6	23.0	33.1	27.4	82.7	27.8	81.4	61.6	27.2	79.0	30.5	41.4	13.4	29.2	14.9	43.2
HRDA [23]	One	86.7	22.0	81.2	26.8	25.8	30.2	40.4	33.6	84.8	24.3	77.8	63.2	32.3	84.7	31.1	40.6	19.4	26.5	14.0	44.5
CACDA * [16]	One	80.9	32.6	85.8	36.1	30.7	40.7	43.7	41.7	84.1	30.7	84.5	65.4	27.6	86.0	36.5	51.4	24.1	26.7	30.7	49.5
<b>DAFormer + DATUM</b>	One	88.1	32.8	84.3	26.6	27.7	32.7	35.7	34.9	86.2	36.2	87.6	65.8	35.8	80.2	39.5	44.1	17.1	42.7	43.8	49.6
<b>HRDA + DATUM</b>	One	82.1	31.9	80.9	21.3	27.6	38.6	43.5	41.0	87.1	33.1	87.2	70.8	37.5	71.3	38.2	48.5	22.9	44.2	54.0	50.6
Encoder: MiT-B5																					
Source-Only ◊	None	-	-																		45.6
DAFormer ◊ [22]	All	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
HRDA ◊ [23]	All	96.4	74.4	91.0	61.6	51.5	57.1	63.9	69.3	91.3	48.4	94.2	79.0	52.9	93.9	84.1	85.7	75.9	63.9	67.5	73.8
DAFormer [22]	One	84.0	18.2	83.0	35.6	20.0	33.8	40.1	35.3	86.4	37.4	82.7	66.6	31.6	85.0	37.0	40.6	36.7	33.3	29.0	48.2
HRDA [23]	One	84.3	29.2	84.3	44.3	23.2	43.9	48.7	39.2	88.2	41.0	82.6	70.5	36.9	85.8	43.7	51.4	42.2	34.7	30.6	52.9
CACDA * [16]	One	83.4	35.3	87.1	44.8	32.3	42.5	50.2	52.5	88.0	46.1	90.4	66.7	25.6	88.6	50.3	50.8	44.5	34.4	38.6	55.4
<b>DAFormer + DATUM</b>	One	86.2	29.0	87.1	41.5	35.6	44.7	48.5	42.7	88.4	42.4	88.3	68.8	35.9	89.7	57.1	57.6	27.8	46.8	53.2	56.4
<b>HRDA + DATUM</b>	One	87.1	32.0	88.2	49.6	40.4	49.5	54.8	43.6	89.9	44.6	91.3	74.9	45.7	91.4	61.7	67.0	37.1	57.7	55.8	61.2

**thetic** target dataset generated by DATUM, alongside source. For a fair comparison with baselines, we use the DAFormer network architecture (with MiT-B5 backbone), which has demonstrated superior effectiveness compared to weaker counterparts, such as ResNet-101 [22]. However, as performance metrics for some older OSUDA methods [3, 29] are not available with a DAFormer-like architecture, we also experiment using DeepLabV2 with ResNet-101 backbone.

**Main results.** In Tab. 1 and Tab. 2 we report the results on the GTA → Cityscapes and the SYNTHIA → Cityscapes benchmarks, respectively, under the traditional UDA as well as the OSUDA setting. The traditional UDA setting [22] is denoted as All, as it uses all target samples, while the OSUDA setting is denoted as One since we have access to only a single datum. Following the standard practices from the one/few shot learning literature, we report our results averaged over 3 independent runs using randomly sampled unlabeled real target datum. Also note that in our experiments we report the model performance after the last training iteration, instead

of picking the maximum mIoU.

From the Tab. 1 we notice that using our generated target dataset for training the state-of-the-art UDA methods in the OSUDA setting, greatly improves their performances, independent of the backbone. For instance, DAFormer + DATUM is +6.4% better (43.2 → 49.6%) than DAFormer, with the ResNet-101 as backbone. Similar trends can be noticed when using the MiT-B5 backbone, where we improve HRDA by +8.3% (i.e., from 52.9% → 61.2%). Overall, for the GTA → Cityscapes with the MiT-B5 as backbone, we beat the best competitor CACDA [16] by significant margins (55.4% versus our 61.2%). Interestingly, we observe that while using the ResNet-101 backbone, our data generation can even outperform UDA methods that use all the original target dataset, e.g., CycleGAN and ASM.

From Tab. 2, which reports the performance on the SYNTHIA → Cityscapes benchmark, we observe similar results. Pairing our generated dataset with UDA methods consistently improves performance under the OSUDA setting.

Table 2. Comparison with state-of-the-art methods for UDA and OSUDA on the GTA → Cityscapes benchmark. #TS denotes the number of *real* target samples used during training, which are color coded as None, All and One. Methods using ResNet-101 [17] and MiT-B5 [56] are shown in the top and bottom halves, respectively. As an example, DaFormer + DATUM denotes DAFormer trained using the synthetic images generated by our DATUM. mIoU<sup>13</sup> and mIoU<sup>16</sup> denote the mIoU computed using the 13 and 16 classes, respectively [47, 48, 52]. \*: results from CACDA [16]; ◊: results from HRDA [23]; and †: results from ASM [29].

	#TS	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Sky	Person	Rider	Car	Bus	M.bike	Bike	mIoU <sup>16</sup>	mIoU <sup>13</sup>
Encoder: ResNet-101																			
Source-Only *	None	36.3	14.6	68.8	9.2	0.2	24.4	5.6	9.1	69.0	79.4	52.5	11.3	49.8	9.5	11.0	20.7	29.5	33.7
AdaptSeg † [48]	All	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
CBST † [62]	All	53.6	23.7	75.0	-	-	-	23.5	26.3	84.8	74.7	67.2	17.5	84.5	28.4	15.2	55.8	-	48.4
DACS ◊ [47]	All	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	90.8	67.6	38.3	82.9	38.9	28.5	47.6	48.3	54.8
ProDA ◊ [59]	All	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	84.4	74.2	24.3	88.2	51.1	40.5	45.6	55.5	62.0
DAFormer ◊ [22]	All	71.5	30.4	85.4	26.2	3.4	40.9	45.9	52.3	84.3	81.4	69.7	42.7	86.9	52.5	49.3	59.4	55.1	61.9
HRDA ◊ [23]	All	85.8	47.3	87.3	27.3	1.4	50.5	57.8	61.0	87.4	89.1	76.2	48.5	87.3	49.3	55.0	68.2	61.2	69.2
CBST † [62]	One	59.6	24.1	72.9	-	-	-	5.5	13.8	72.2	69.8	55.3	21.1	57.1	17.4	13.8	18.5	-	38.5
AdaptSeg † [48]	One	64.1	25.6	75.3	-	-	-	4.7	2.7	77.0	70.0	52.2	20.6	51.3	22.4	19.9	22.3	-	39.1
OST † [3]	One	75.3	31.6	72.1	-	-	-	12.3	9.3	76.1	71.1	51.1	17.7	68.9	19.0	26.3	25.4	-	42.8
ASM † [29]	One	73.5	29.0	75.2	-	-	-	10.9	10.1	78.1	73.2	56.0	23.7	76.9	23.3	24.7	18.2	-	44.1
SMPPM * [55]	One	79.3	35.3	75.9	5.6	16.6	29.8	25.4	22.7	79.9	76.8	54.6	23.5	60.2	23.9	21.2	36.6	41.4	47.3
DAFormer [22]	One	69.3	26.3	76.3	5.8	0.5	28.5	16.7	24.9	73.7	74.9	59.5	28.5	74.5	28.0	21.8	44.6	40.9	47.1
HRDA [23]	One	61.0	24.1	76.7	7.5	0.3	34.5	21.8	29.2	77.4	78.9	64.2	28.5	77.1	25.0	29.8	43.4	42.5	48.5
CACDA * [16]	One	82.5	33.8	77.8	12.6	0.8	34.2	30.8	34.4	79.8	82.4	55.4	30.7	72.5	28.4	15.9	47.8	45.0	51.7
<b>DAFormer + DATUM</b>	One	79.3	32.9	80.6	17.7	0.4	32.4	22.2	36.9	82.4	81.6	65.7	36.0	76.2	26.0	31.2	50.3	47.0	53.5
<b>HRDA + DATUM</b>	One	86.5	39.3	83.2	17.9	0.2	42.8	24.0	45.1	84.1	85.9	72.7	39.2	86.1	31.4	44.5	56.7	52.5	59.4
Encoder: MiT-B5																			
DAFormer ◊ [22]	All	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	89.8	73.2	48.2	87.2	53.2	53.9	61.7	60.9	67.4
HRDA ◊ [23]	All	85.2	47.7	88.8	49.5	4.8	57.2	65.7	60.9	85.3	92.9	79.4	52.8	89.0	64.7	63.9	64.9	65.8	72.4
DAFormer [22]	One	71.2	25.7	82.3	20.5	0.9	37.0	30.0	28.5	83.7	86.8	61.2	31.0	73.3	24.8	14.1	28.8	43.7	48.9
HRDA [23]	One	73.2	27.6	81.8	24.0	0.5	43.5	42.0	32.5	85.3	87.2	65.3	30.3	74.5	29.8	13.4	42.6	47.1	52.3
CACDA * [16]	One	81.4	37.3	84.8	19.5	1.2	43.7	43.0	34.4	86.5	90.0	63.8	32.8	79.6	42.7	28.0	47.2	51.0	57.8
<b>DAFormer + DATUM</b>	One	79.6	28.8	85.6	30.9	1.4	45.6	43.0	46.5	85.9	89.7	70.3	38.4	84.8	56.0	39.5	52.1	54.9	61.1
<b>HRDA + DATUM</b>	One	83.2	31.8	86.6	37.4	0.8	51.4	46.9	52.0	87.8	92.0	76.1	43.7	88.4	56.3	48.5	57.1	58.7	64.9

Compared to the best competitor method CACDA, DATUM helps achieve the new state-of-the-art results, by comprehensively outperforming CACDA by +2.0% and +3.1% in mIoU<sup>13</sup>. We believe that these findings are highly significant in bridging the gap between OSUDA and standard UDA.

Table 3. Comparison with style transfer-based OSUDA methods

	ResNet-101	MiT-B5
RAIN [29]	42.7	53.4
PT+CDIR [16]	48.5	54.0
<b>DATUM (Ours)</b>	<b>50.6</b>	<b>57.2</b>

**Comparison with style-transfer methods.** Given that DATUM is akin to data augmentation in image stylization, we compare it against two style transfer techniques used in existing OSUDA methods: RAIN [29] and PT+CDIR [16]. Tab. 3 reports the results. We observe that generating novel scenes with DATUM is more impactful than simply augmenting the source images with the target style as in the other methods.

### 4.3. Ablation analysis

To examine the effectiveness of DATUM, in this section we conduct thorough ablation analyses of each component associated with it. All ablations are carried out on GTA → Cityscapes benchmark with DAFormer [22] using only one random datum from the target domain.

**Impact of number of shots.** To investigate the impact of the number of real target samples (or #TS) on the OSUDA performance, we conduct an ablation study where we vary the #TS and personalize SD with DATUM for a varied number of training iterations. In Fig. 4 we plot the performance of DAFormer + DATUM for different #TS and compare it with SD. We observe that for lower #TS (*i.e.*, one shot) DATUM achieves the best performance, and the mIoU gradually degrades with prolonged training. This is because the SD overfits on the single target image and loses its ability to generate diverse scenes. The issue is less severe when the #TS increases to 10 (*i.e.*, ten shot), and the mIoU is fairly stable. Nevertheless, DATUM generates more informative target

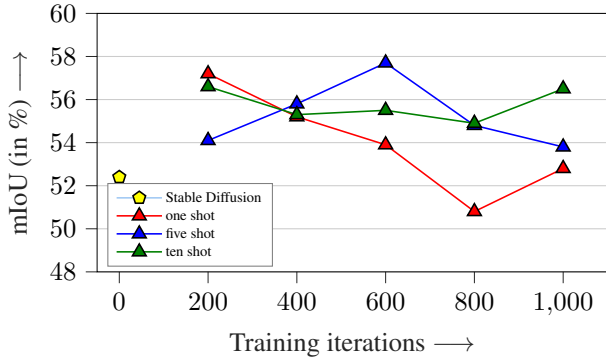


Figure 4. Impact of number of shots (#TS) on the mIoU (in %)

Table 4. Impact of training and inference prompts on the mIoU.

Training prompt	Inference prompt	classes	mIoU
"a photo of a $V_*$ urban scene"	"a photo of a $V_*$ urban scene"	-	52.9
	"a photo of a $V_*$ [CLS]"	things	<b>57.2</b>
	"a photo of a $V_*$ [CLS]"	things + stuff	56.7
	"a photo of a $V_*$ [CLS] seen from the dash cam"	things	55.5
"a photo of a $V_*$ scene from a car"	"a photo of $V_*$ scene from a car"	things	53.0
	"a photo of a $V_*$ [CLS]"	things	56.8
	"a photo of [CLS] in a $V_*$ scene from a car"	things	55.4

images than SD, highlighting the need for incorporating the target style into the synthetic dataset generation process.

**Impact of prompts.** Since DATUM depends on the choice of prompts used during training and inference, here we ablate the impact of training and inference prompts by quantitatively measuring the mIoU for different prompts and report the results in Tab. 4. We observe that the combination of training prompt  $p = \text{"a photo of a } V_* \text{ urban scene"}$  and class-aware inference prompt  $p = \text{"a photo of a } V_* \text{ [CLS]"}$  leads to the best results (second row). When compared to the class agnostic inference prompt  $p = \text{"a photo of a } V_* \text{ urban scene"}$  (first row), the performance increases by +4.3%. This demonstrates that grounding DATUM with things/objects of interest leads to more meaningful scene composition, and provides more information to the segmentation model. Using the *stuff* classes (e.g., *sky*, *building*) in the inference prompts (third row) results in a slightly lower performance compared to using only *things* classes (second row).

Given that the target dataset Cityscapes is captured with sensors mounted on a car, we make an attempt to tailor the inference prompts for such a use-case. Specifically, we use the inference prompt  $p = \text{"a photo of a } V_* \text{ seen from the dash cam"}$ . We notice that usage of such prompt does not bring any improvement, and rather leads to worsened performance.

Next, we make the training prompt more suited for a driving scenario by using  $p = \text{"a photo of a } V_* \text{ scene from a car"}$  and experiment with some inference prompts that are essentially nuanced variations of the training prompt. The results are reported in the lower part of Tab. 4. We observe

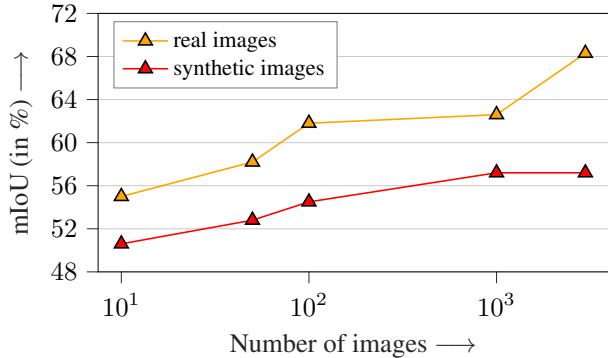


Figure 5. Impact of the cardinality of the generated target dataset on the mIoU in the one-shot setting. It is compared with adaptation on the real data

that adding the phrase *"scene from a car"* to the training prompt has no positive impact in the training of DATUM. It is worth noting that the retention of the *prior preservation* loss caused our best result to decrease from 57.2% to 54.8%.

**Impact of generated dataset cardinality.** Here, we examine the impact of the cardinality of the target dataset generated by DATUM using a single real target image (i.e., one-shot) on the performance of the segmentation model. In Fig. 5 we plot the mIoU from DAFormer versus the generated dataset size and also compare with training on the real target dataset of the same cardinality. We observe that having the same quantity of real target samples leads to better performance with respect to purely synthetic data. This is expected as real data always contains more targeted information than synthetic data. However, one must appreciate the fact that having 1000 synthetic data leads to a better performance than having 10 real samples, which can be difficult to collect in some applications. Thus, our DATUM is most effective when working with a very small budget of real target data.

## 5. Conclusions

We proposed a synthetic data generation method DATUM for the task of one-shot unsupervised domain adaptation, that uses a single image from the target domain to personalize a pre-trained text-to-image diffusion model. The personalization leads to a synthetic target dataset that faithfully depicts the style and content of the target domain, whereas the text-conditioning ability allows for generating diverse scenes with desired semantic objects. When pairing DATUM with modern UDA methods, it outperforms all state-of-the-art OSUDA methods, thus paving the path for future research in this few-shot learning paradigm.

**Acknowledgements.** This paper has been supported by the French National Research Agency (ANR) in the framework of its JCJC This work was granted access to the HPC resources of IDRIS under the allocation AD011013071 made by GENCI.



## References

- [1] Mohamed Akrouf, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincsó, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. *arXiv preprint arXiv:2301.04802*, 2023. [3](#)
- [2] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021. [2](#)
- [3] Sagie Benaïm and Lior Wolf. One-shot unsupervised cross domain translation. *advances in neural information processing systems*, 31, 2018. [1](#), [2](#), [5](#), [6](#), [7](#)
- [4] Róger Bermúdez-Chacón, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. A domain-adaptive two-stream u-net for electron microscopy image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 400–404. IEEE, 2018. [2](#)
- [5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocomp: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. [5](#)
- [6] Senay Cakir, Marcel Gauß, Kai Häppeler, Yassine Ounajjar, Fabian Heinle, and Reiner Marchthaler. Semantic segmentation for autonomous driving: Model evaluation, dataset generation, perspective comparison, and real-time capability. *arXiv preprint arXiv:2207.12939*, 2022. [1](#)
- [7] Bi-ke Chen, Chen Gong, and Jian Yang. Importance-aware semantic segmentation for autonomous driving system. In *IJCAI*, pages 1504–1510, 2017. [1](#)
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. [1](#), [5](#)
- [9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. [5](#)
- [10] Gabriela Csurka, Riccardo Volpi, and Boris Chidlovskii. Unsupervised domain adaptation for semantic image segmentation: a comprehensive survey. *arXiv preprint arXiv:2112.03241*, 2021. [1](#)
- [11] Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 14(1-2):1–162, 2022. [1](#)
- [12] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018. [2](#)
- [13] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. [2](#)
- [14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Representation Learning*, 2023. [2](#), [3](#)
- [15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, 2016. [2](#)
- [16] Rui Gong, Qin Wang, Dengxin Dai, and Luc Van Gool. One-shot domain adaptive and generalizable semantic segmentation with class-aware cross-domain transformers. *arXiv preprint arXiv:2212.07292*, 2022. [1](#), [2](#), [5](#), [6](#), [7](#)
- [17] Kaifeng He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [5](#), [6](#), [7](#)
- [18] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Representation Learning*, 2023. [3](#)
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [2](#), [3](#)
- [20] 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 *International Conference on Machine Learning*, pages 1989–1998. Pmlr, 2018. [1](#), [2](#), [3](#)
- [21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. [2](#)
- [22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. [2](#), [5](#), [6](#), [7](#)
- [23] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. [2](#), [5](#), [6](#), [7](#)
- [24] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 590–605, 2018. [2](#)
- [25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. [2](#)
- [26] Juana Valeria Hurtado and Abhinav Valada. Semantic scene segmentation for robotics. In *Deep Learning for Robot Perception and Cognition*, pages 279–311. Elsevier, 2022. [1](#)
- [27] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. [2](#)
- [28] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with

- composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. [2](#)
- [29] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 33:20612–20623, 2020. [1](#), [2](#), [5](#), [6](#), [7](#)
- [30] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. [2](#)
- [31] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in neural information processing systems*, 30, 2017. [2](#)
- [32] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020. [2](#)
- [33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. [3](#)
- [34] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021. [1](#)
- [35] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. [5](#)
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. [1](#), [2](#), [3](#), [5](#)
- [37] Robin Rombach and Patrick Esser. Compvis/stable-diffusion-v1-4. <https://huggingface.co/CompVis/stable-diffusion-v1-4>. Accessed on March 20, 2023. [5](#)
- [38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. [3](#)
- [39] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016. [5](#)
- [40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. [2](#), [3](#), [4](#), [5](#)
- [41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. [2](#), [3](#)
- [42] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning (s) from a synthetic imagenet clone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. [3](#)
- [43] Tong Shen, Dong Gong, Wei Zhang, Chunhua Shen, and Tao Mei. Regularizing proxies with multi-adversarial training for unsupervised domain-adaptive semantic segmentation. *arXiv preprint arXiv:1907.12282*, 2019. [2](#)
- [44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. [3](#)
- [45] Marco Toldo, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for mobile semantic segmentation based on cycle consistency and feature alignment. *Image and Vision Computing*, 95:103889, 2020. [2](#)
- [46] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. [1](#)
- [47] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. [1](#), [5](#), [6](#), [7](#)
- [48] 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. [2](#), [5](#), [6](#), [7](#)
- [49] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. [2](#)
- [50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [3](#)
- [51] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. [5](#)

- [52] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 1, 2, 3, 5, 7
- [53] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1
- [54] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7173–7182, 2019. 2
- [55] Xinyi Wu, Zhenyao Wu, Yuhang Lu, Lili Ju, and Song Wang. Style mixing and patchwise prototypical matching for one-shot unsupervised domain adaptive semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2740–2749, 2022. 1, 2, 5, 6, 7
- [56] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1, 5, 6, 7
- [57] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 2
- [58] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3
- [59] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 5, 6, 7
- [60] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021. 2
- [61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 5, 6
- [62] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 6, 7