Bridging Domains with Minimal Supervision: Domain Adaptation and Generalization for Semantic Segmentation

Yasser Benigmim (3rd year PhD student)

Multimedia Team (LTCI, Telecom Paris) & VISTA Team (LIX, Ecole Polytechnique)

Supervisors:

Stéphane Lathuilière (LTCI, Telecom Paris)
Vicky Kalogeiton (LIX, Ecole Polytechnique)
Slim Essid (LTCI, Telecom Paris)

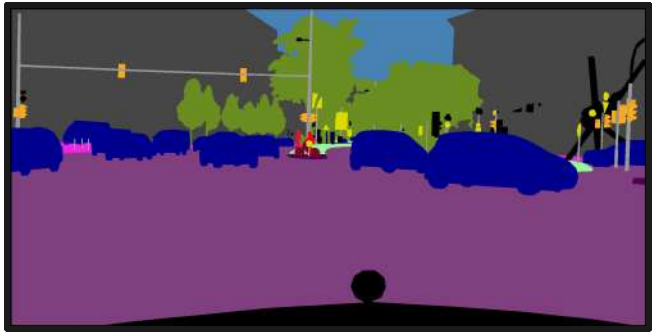
Task : Semantic Segmentation

- The objective of Semantic Segmentation is to assign a class for every pixel in the image

- A real-life HR image (2048x1024) contains ~ 2×10^6 pixels

- It takes around ~90min to manually segment one image

- Training on huge amounts of labelled real-life data for Semantic Segmentation is very **expensive**



Task : Semantic Segmentation

To alleviate the problem of annotations, multiple research axes have been proposed :

- Weakly-Supervised Learning, Semi-Supervised Learning …
- **Train a model in a supervised way on a dataset easy to collect like a synthetic one -> Use the model at inference on a real life dataset**

→ **Domain shift !**

Task : Semantic Segmentation

Domain shift

Synthetic data (GTA5)

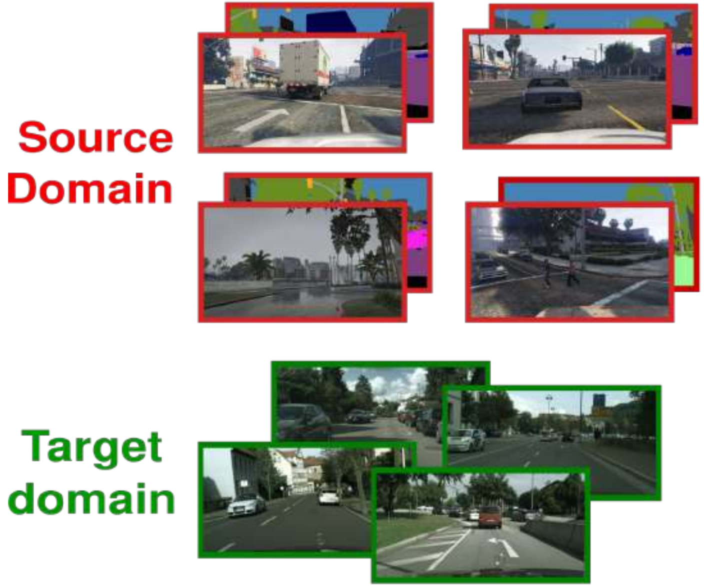
real-life data (Cityscapes)

Image credits : S.Richter, et al. "Playing for Data: Ground Truth from Computer Games" ECCV 2016
M. Cordts, et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding” CVPR 2016

Setting : Unsupervised Domain Adaptation (UDA)

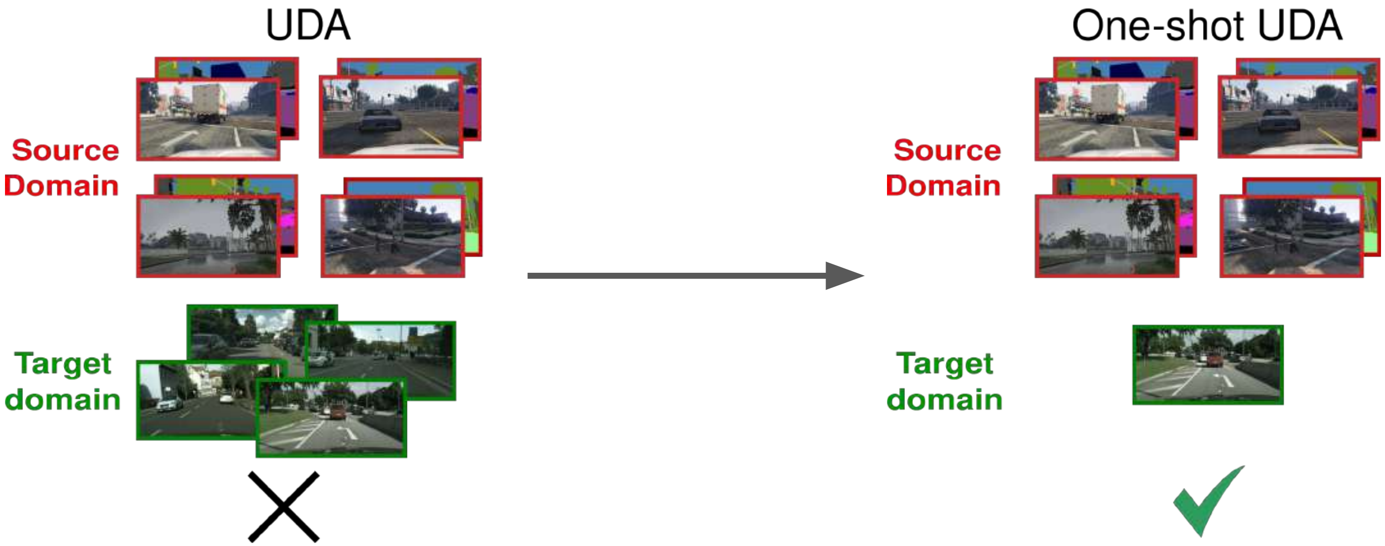
- UDA assumes having access during training to **labelled data from source domain (easy to obtain)** and to **unlabelled data from target domain**

- During test, we deploy the model on **unseen images from target domain**



Setting : One-shot UDA

- We have access to a **labeled source domain** and **one unlabeled image** from target domain



DATUM : One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models

Yasser Benigmim [1,2], Subhankar Roy [1], Slim Essid [1], Vicky Kalogeiton [2], Stéphane Lathuilière [1]

[1] LTCI, Télécom-Paris, Institut Polytechnique de Paris

[2] LIX, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris

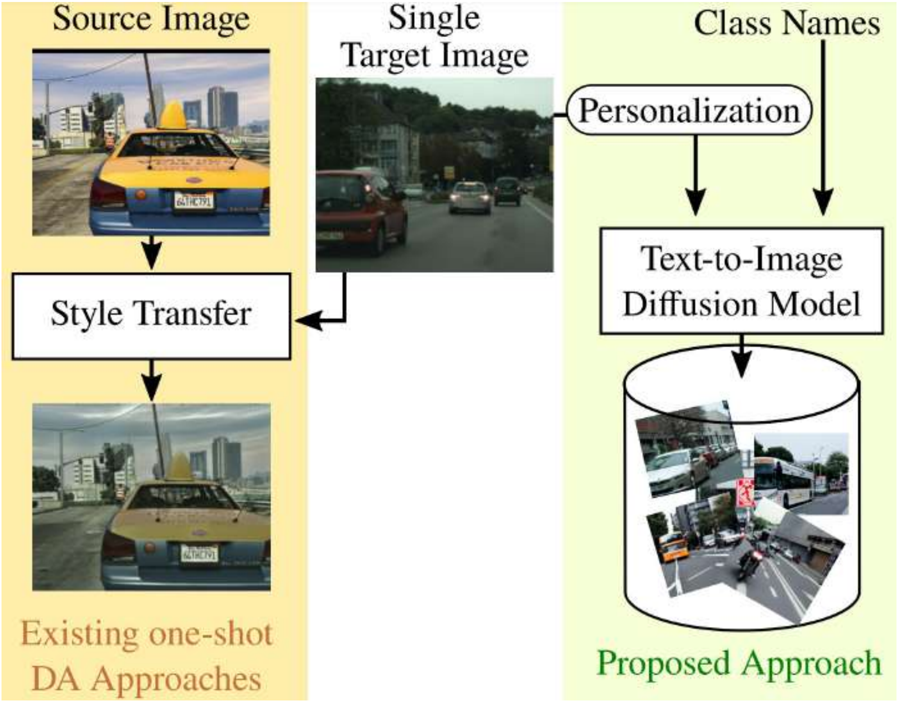
https://github.com/yasserben/DATUM

CVPR-W’23 (Generative Models for Computer Vision)

Previous works

- Previous works [1,2], rely on **style transfer** to adapt the source images to the target and train on the stylized images using original GT labels

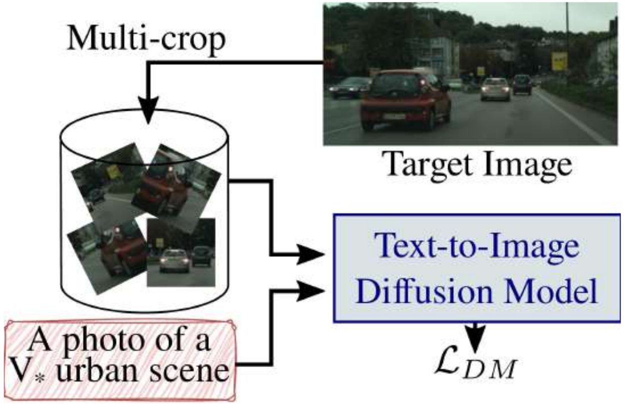
- Our method uses a T2I diffusion model to generate a pseudo-target domain then trains any UDA method on it.



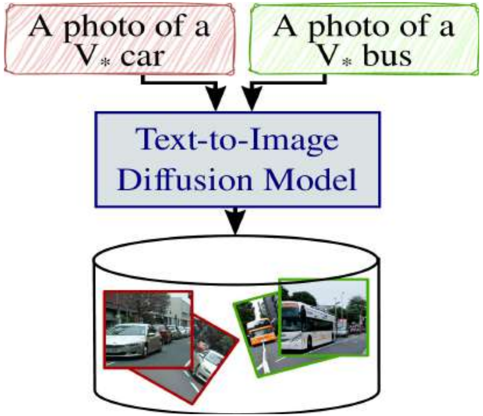
1 : Y.Luo, et al. "Adversarial style mining for one-shot unsupervised domain adaptation." NeurIPS 2020
2: X.Wu, et al. "Style mixing and patchwise prototypical matching for one-shot unsupervised domain adaptive semantic segmentation." AAAI 2022

Method : DATUM

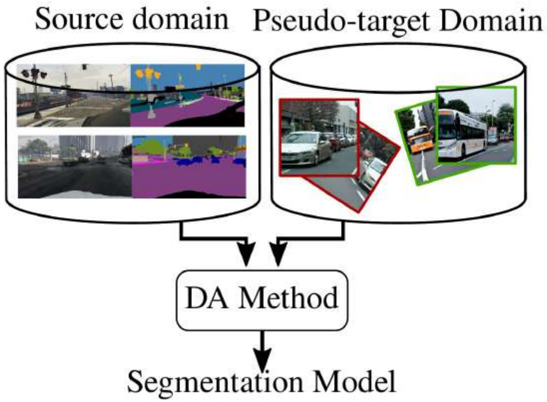
- DATUM is composed of three steps : **Personalization, Generation and Adaptation**



1. Personalization



2. Generation



3. Adaptation

Previous work : Dreambooth

- Dreambooth is a method that allows the user to **personalize** a text-to-image diffusion model
- The key idea behind Dreambooth is to **associate a unique identifier to the concept** we want to inject in a diffusion model

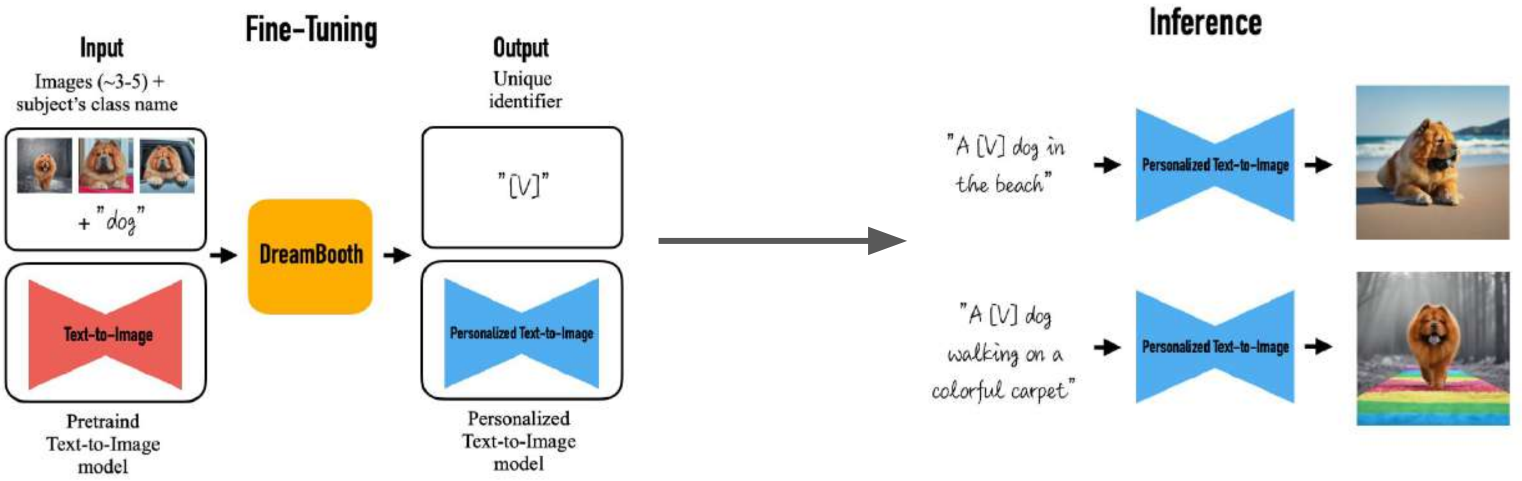
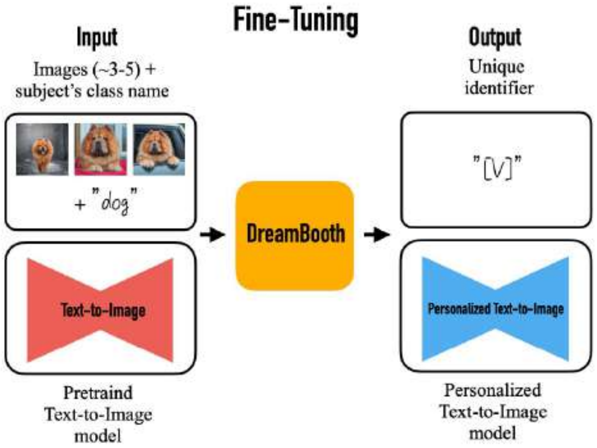


Image credits : Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." CVPR 2023

Step 1 : Personalization

- We **finetune** our diffusion model with the single target image using **Dreambooth**



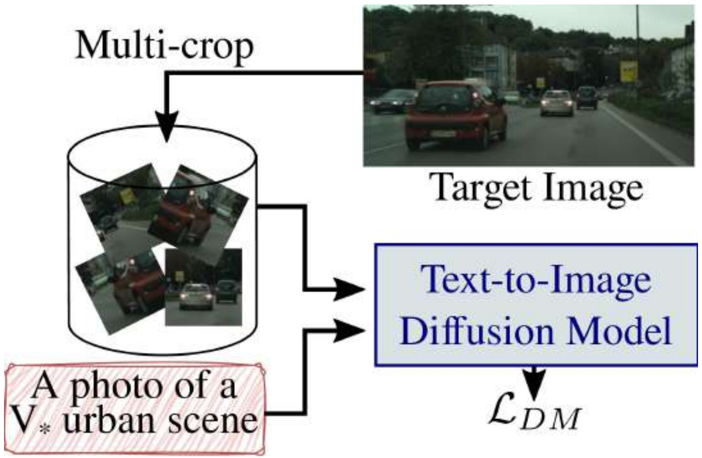
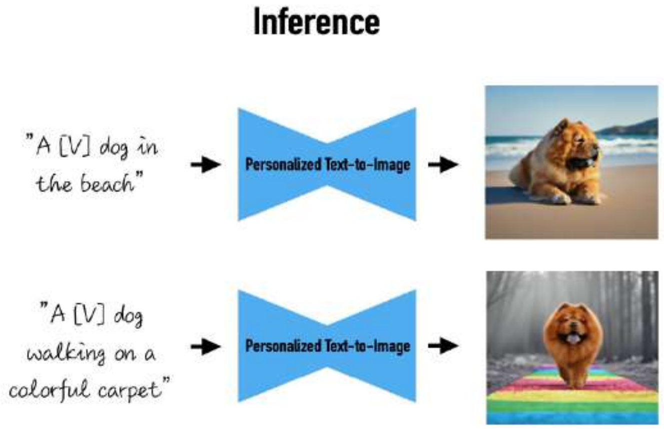


Image credits : Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." CVPR 2023

Step 2 : Generation

- We **generate new images** using the unique identifier associated with the target image
- We use **class-specific prompts + unique identifier** to increase image diversity



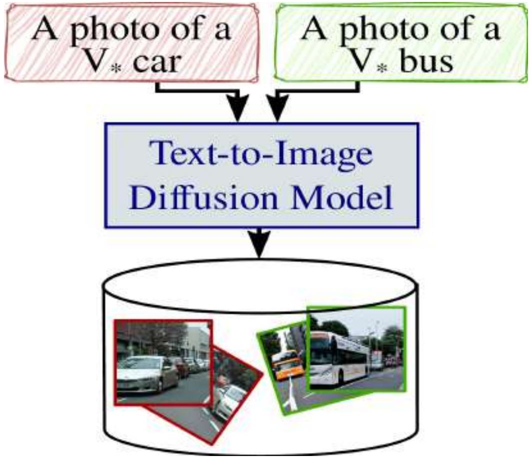


Image credits : Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." CVPR 2023

Visualization

"a photo of a car"

Dreambooth

"a photo of a V car"*

Visualization

"a photo of a bus"

Dreambooth

"a photo of a V* bus"

Visualization

"a photo of a traffic sign"

Dreambooth

"a photo of a V traffic sign"*

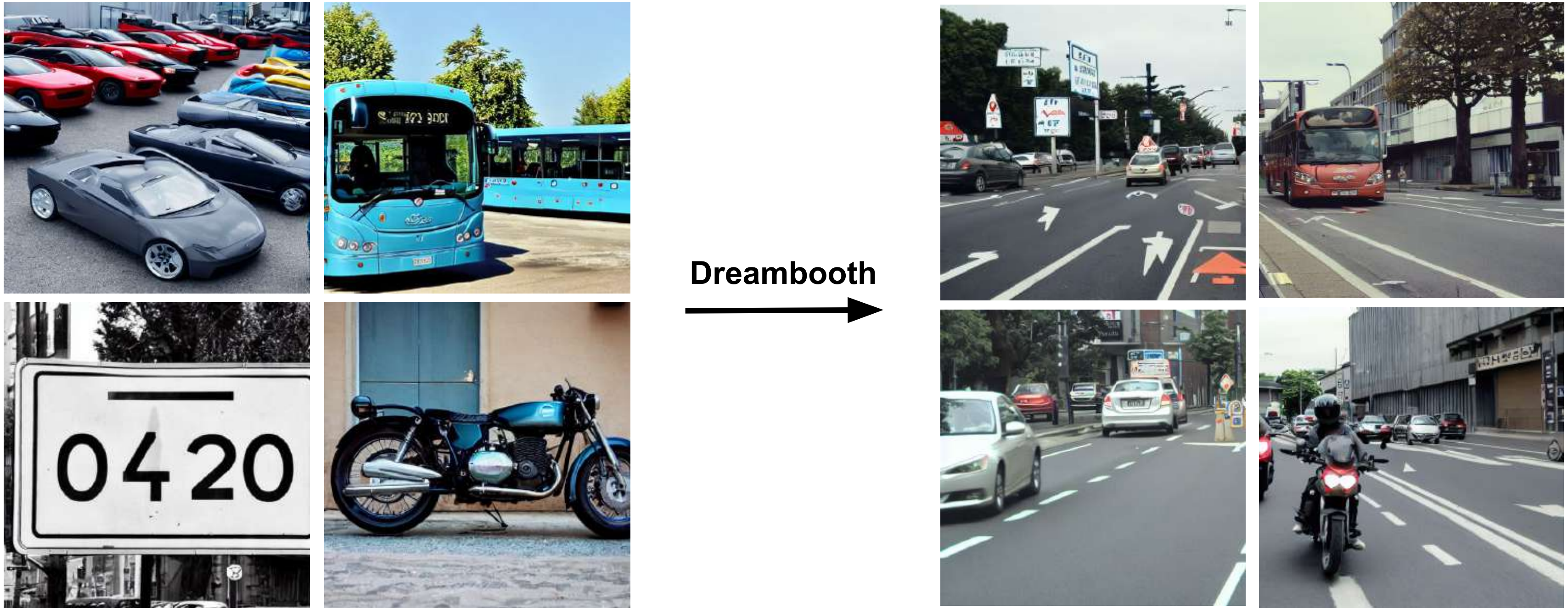
Visualization

“a photo of a motorcycle”

Dreambooth

“a photo of a V* motorcycle”

Visualization



Visualization

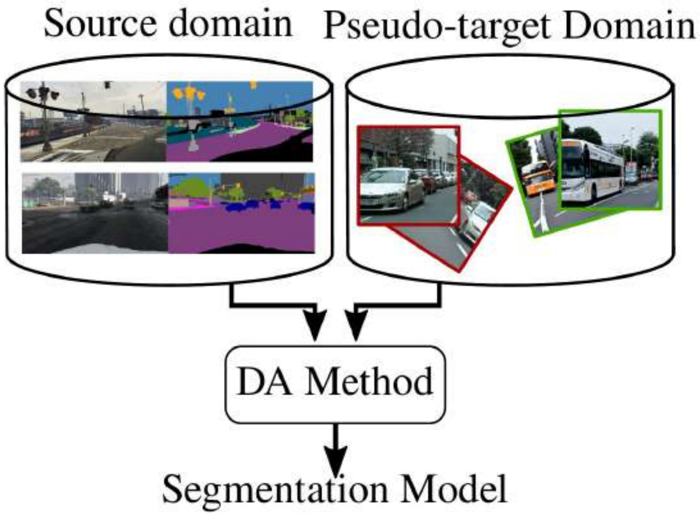
DB

pseudo-target domain

target domain

Step 3 : Adaptation

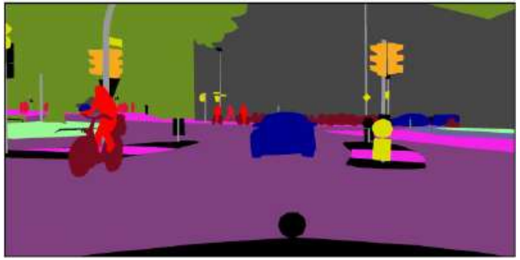
- We can **inject the generated dataset** into any previous UDA framework
- DATUM is **plug-and-play** method making **any UDA method work in a data-scarce scenario**



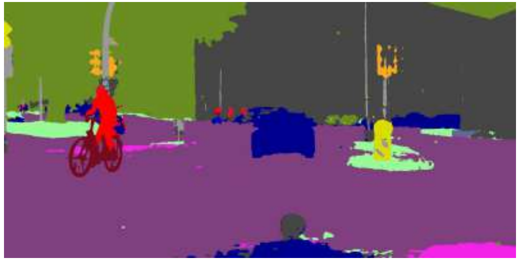
Results

- To evaluate a model on Semantic Segmentation we use the **Intersection over Union** metric

Input image

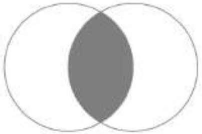


Groundtruth map



predicted map

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



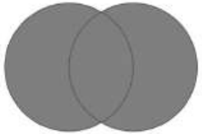
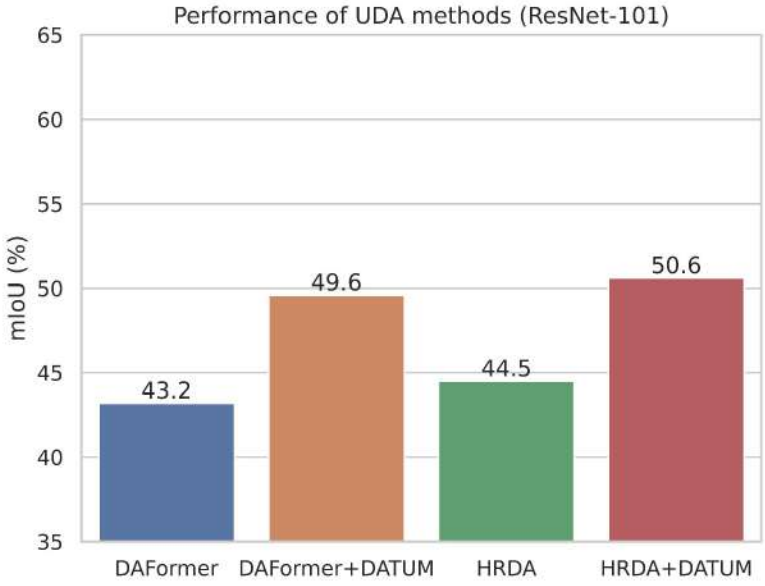


Image credits : https://medium.com/@cyborg.team.nitr/miou-calculation-4875f918f4cb

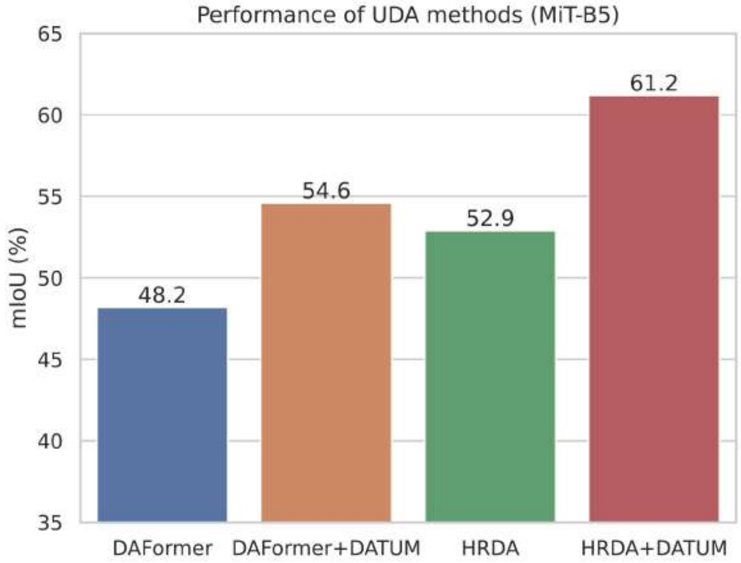
Results

- The performance of **DAFormer (R-101)** in the UDA scenario where all target domain is available (2975 images) is **57.3%**, and **HRDA (R-101)** is **63.0%** on **mIoU**



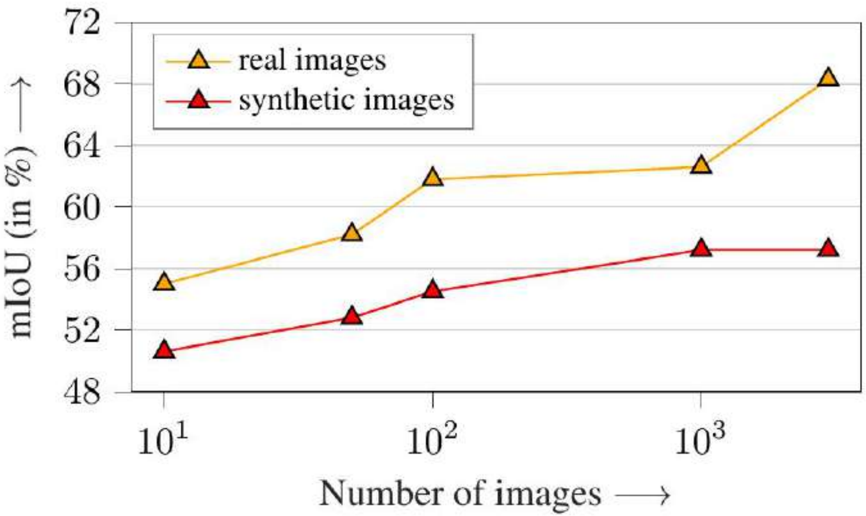
Results

- The performance of **DAFormer (MiT-B5)** in the UDA scenario where all target domain is available (2975 images) is **68.3%**, and **HRDA (MiT-B5)** is **73.8%**



Results

- Increasing the size of the pseudo-target data improves the results
- There remains a gap between the pseudo-target data and real-life one



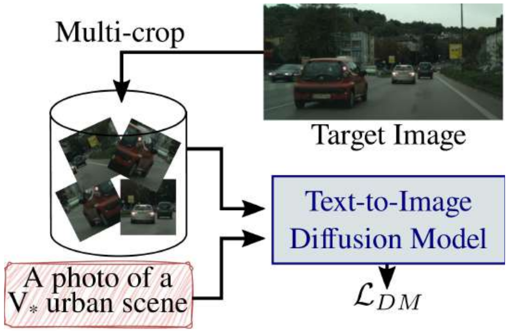
Impact of prompting

- The inference prompt has an impact on the diversity of the generated dataset

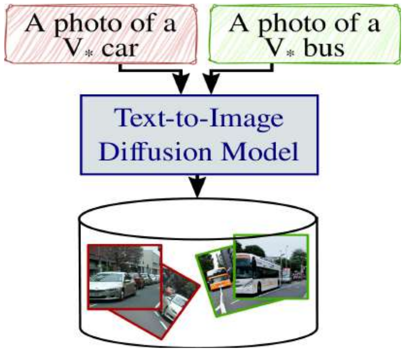
Training prompt	Inference prompt	classes	mIoU
"a photo of a V_* urban scene"	"a photo of a V_* urban scene"	-	52.9
	"a photo of a V_* [CLS]"	things	**57.2**
	"a photo of a V_* [CLS]"	things + stuff	56.7
	"a photo of a V_* [CLS] seen from the dash cam"	things	55.5
"a photo of a V_* scene from a car"	"a 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

"a photo of a V urban scene"*

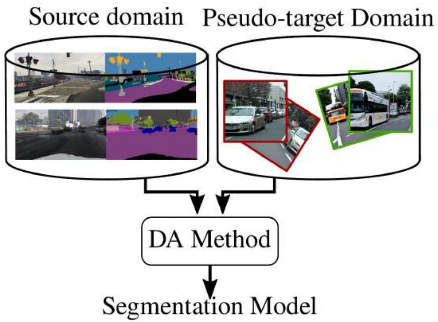
Conclusion



DATUM personalizes T2I diffusion model with a single target image to mimic target domain style



DATUM uses synthetic data for domain adaptation in data-scarce settings.



Integrating DATUM with UDA methods surpasses top OSUDA methods, advancing few-shot learning.

CLOUDS : Collaborating Foundation models for Domain Generalized Semantic Segmentation

Yasser Benigmim [1,2], Subhankar Roy [3], Slim Essid [1], Vicky Kalogeiton [2], Stéphane Lathuilière [1]

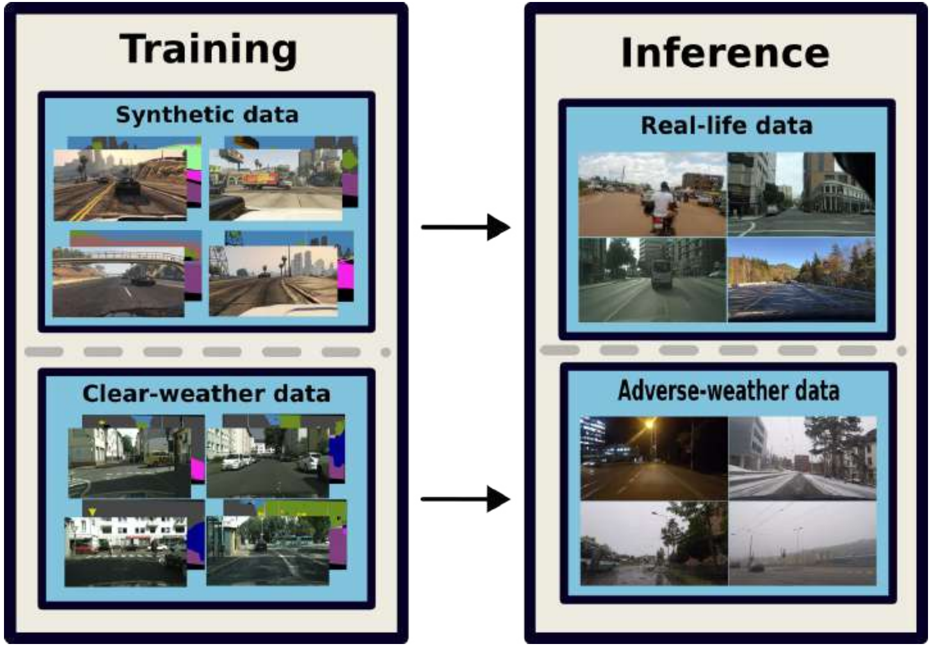
[1] LTCI, Télécom-Paris, Institut Polytechnique de Paris
[2] LIX, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris
[3] University of Aberdeen

https://github.com/yasserben/CLOUDS

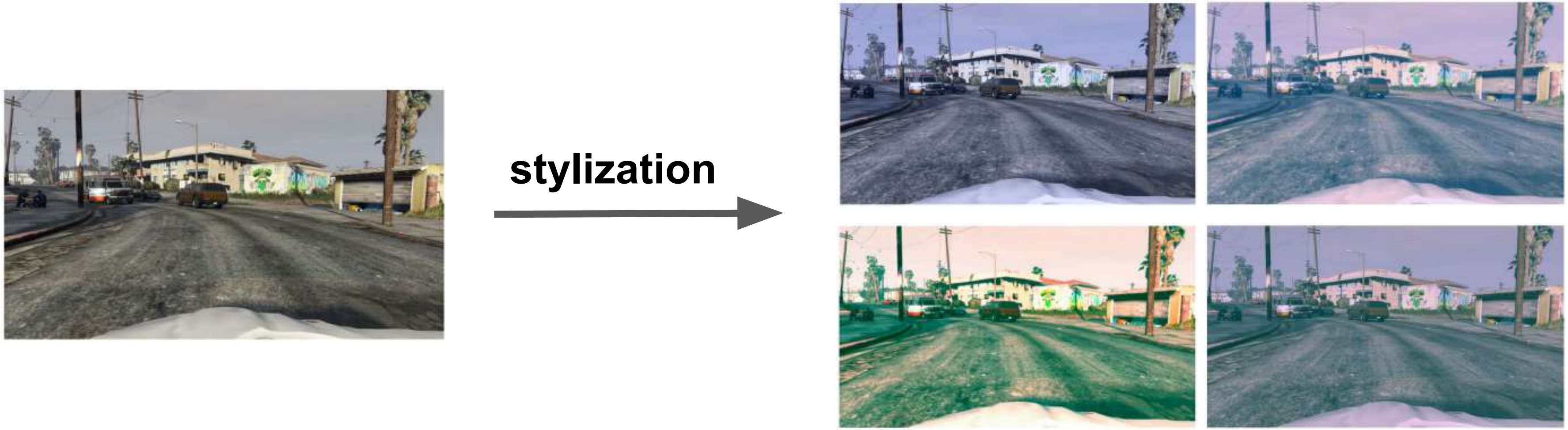
CVPR’24

Domain Generalized Semantic Segmentation (DGSS)

- In DGSS, the model is trained on **labeled source data only** and tested on **unseen domains**



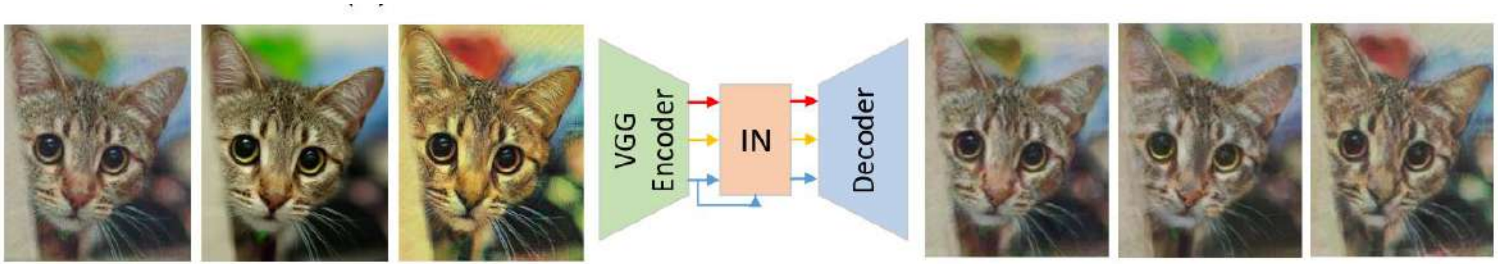
Previous works



Domain Randomization through style diversification only

1 : Z.Zhong, et al. "Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation" NeurIPS 2022

Previous works



Tailor-made modules to eliminate **domain specific features**

1 : X.Pan, et al. "Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net" ECCV 2018

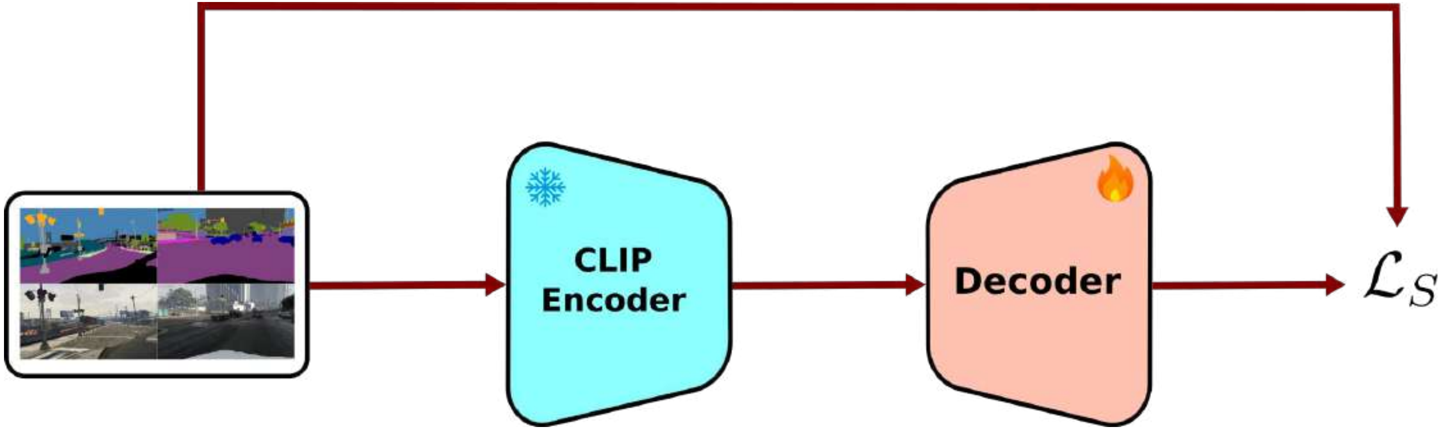
Foundation models

- The rise of large-scale pretrained models, also called Foundation Models (FMs) constitute a new paradigm shift in the field

- We believe that bringing the power of FMs would definitely help advance the setting of DGSS :
    - **Obtain robust feature representations to unseen domains (CLIP)**
    - **Generate diverse images with varied content and styles for self-training (T2I diffusion model + LLM)**
    - **Improve the pseudo labels obtained with self-training (SAM)**

Method

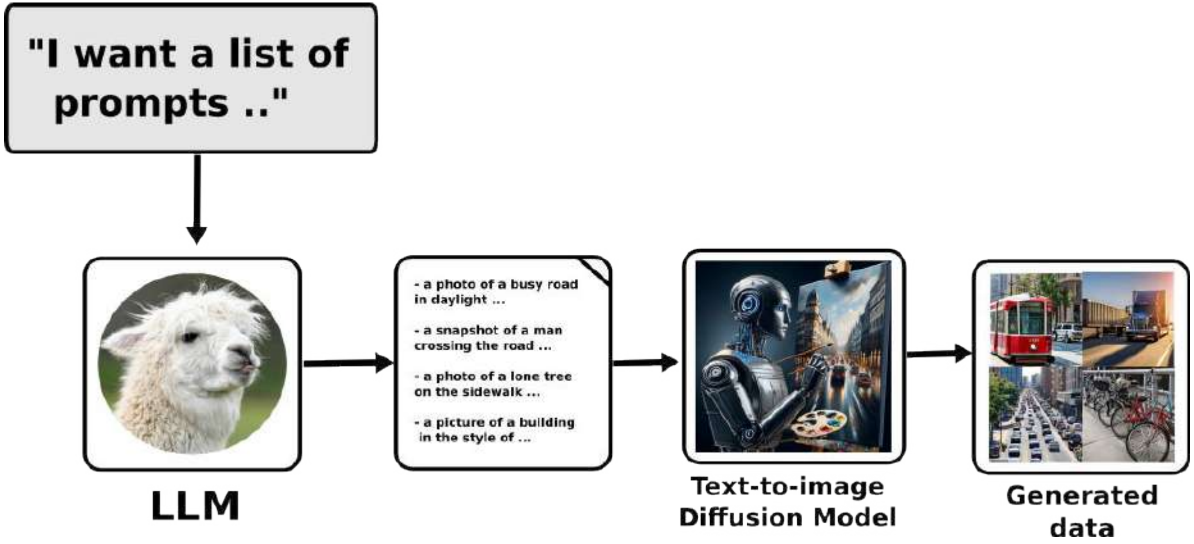
- We use **CLIP** to extract **robust feature representations**
- We freeze the backbone to ensure **preserved generalizability**



Method

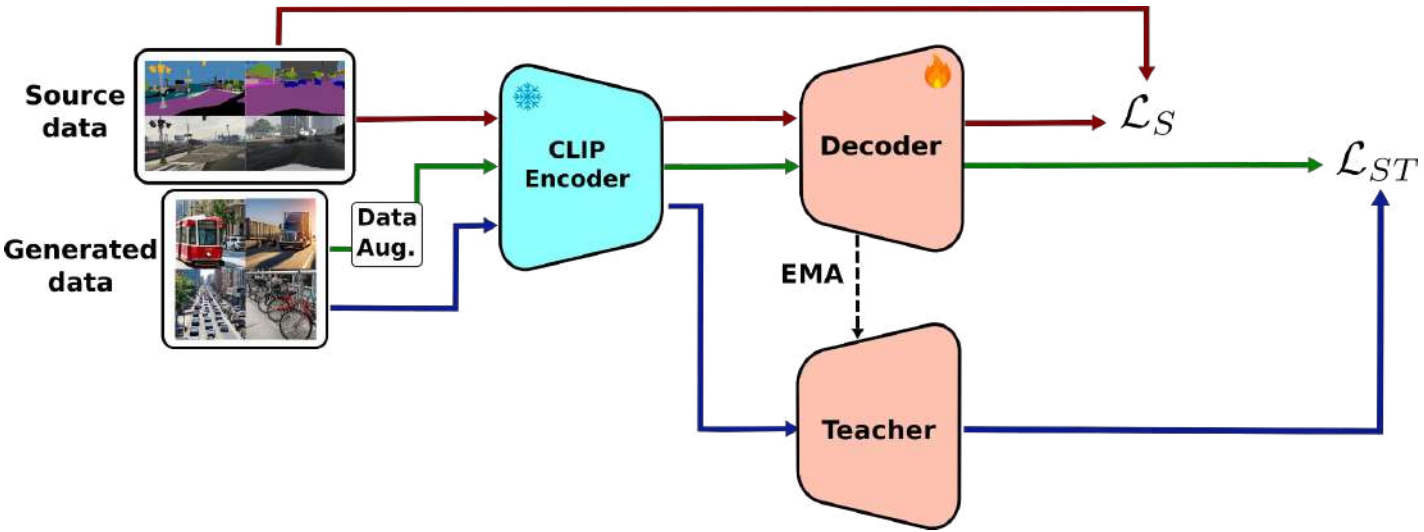
- We generate synthetic data with a **T2I diffusion model** to simulate unseen domains for self-training

- We use an **LLM** to create descriptive text prompts that condition the T2I diffusion model for generating diverse content and styles



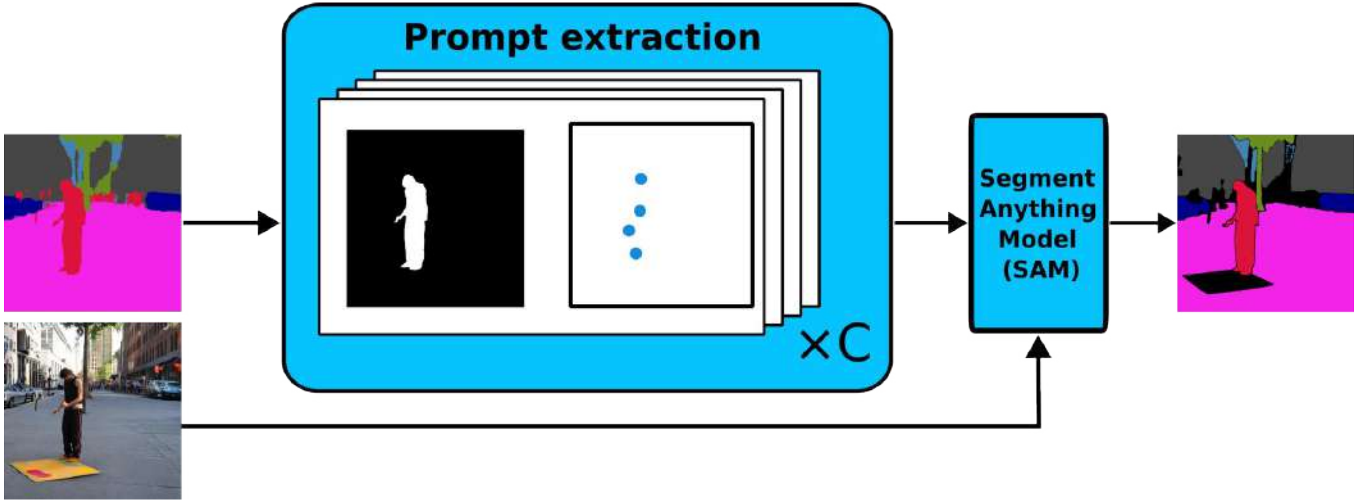
Method

- We **self-train** the model on the generated data using **pseudo-labels (PLs)**

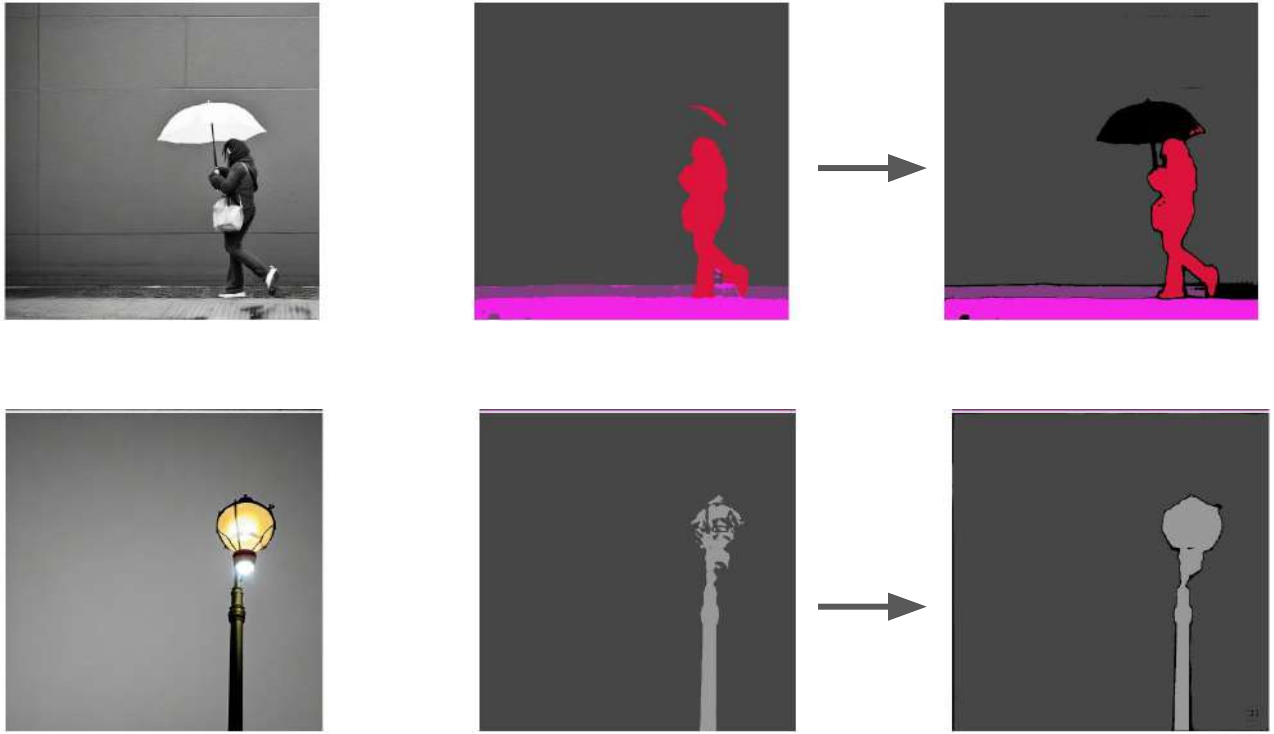


Method

- To improve the **noisy PLs**, we use the **Segment Anything Model (SAM)**
- We extract class-wise masks and point prompts for each noisy PL, feeding them to **SAM** to refine masks



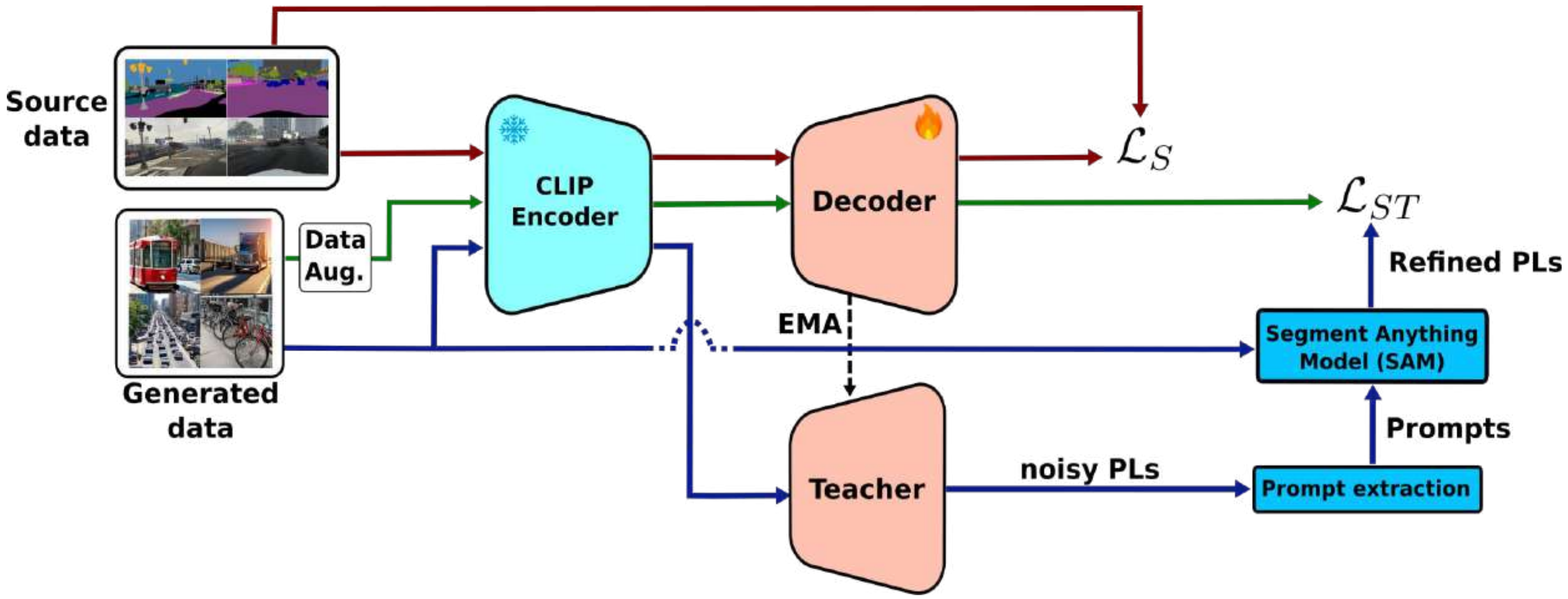
Method



Input image Before After

Method

- We incorporate our PL-refinement module during training for better self-training



Results

- CLOUDS exhibit strong performance on different backbones (ResNet-50, 101 and ConvNext-L)

- on MiT-B5, (pre-trained on ImageNet) CLOUDS outperforms previous SOTA

Method	Encoder	C	B	M	Avg
DRPC [79]	ResNet-50	35.7	31.5	32.7	33.3
SAN-SAW [53]		38.9	35.2	34.5	36.2
MoDify [31]		38.9	33.7	36.2	36.3
TLDR [33]		41.9	34.4	36.8	37.7
CLOUDS (Ours)		**46.1**	**37.6**	**48.1**	**43.9**
DRPC [79]	ResNet-101	37.6	34.4	34.1	35.3
GTR [52]		39.7	35.3	36.4	37.1
FSDR [27]		40.8	37.4	39.6	39.3
SAN-SAW [53]		40.9	36.0	37.3	38.0
TLDR [33]		42.6	35.5	37.5	38.5
HRDA * [26]		34.9	25.0	34.0	31.3
MoDify [31]		43.4	39.5	42.3	41.7
CLOUDS (Ours)		**49.1**	**40.3**	**50.1**	**46.5**
HRDA * [26]	MiT-B5	39.6	32.6	40.0	37.4
CLOUDS (Ours)		**42.2**	**38.3**	**43.6**	**41.4**
FC-CLIP * [78]	ConvNext-L	38.0	29.9	39.0	35.6
CLOUDS (Ours)		**53.4**	**47.0**	**55.8**	**52.1**

Table 2. Comparison with state-of-the-art methods for DGSS on Synthia $\rightarrow$ {Cityscapes (C), BDD (B), Mapillary (M)}. * denotes experiment obtained using the official code

Ablation study

- CLIP alone exhibit strong performance (better than previous SOTA)
- Adding SAM helps to improve the effectiveness of self-training

Backbone	CLIP	{LLM, Diffusion}	SAM	Avg
ResNet-50	✓			50.0
	✓	✓		50.7
	✓	✓	✓	**53.3**
ResNet-101	✓			51.9
	✓	✓		53.3
	✓	✓	✓	**54.7**
ConvNext-L	✓			58.5
	✓	✓		58.6
	✓	✓	✓	**61.5**

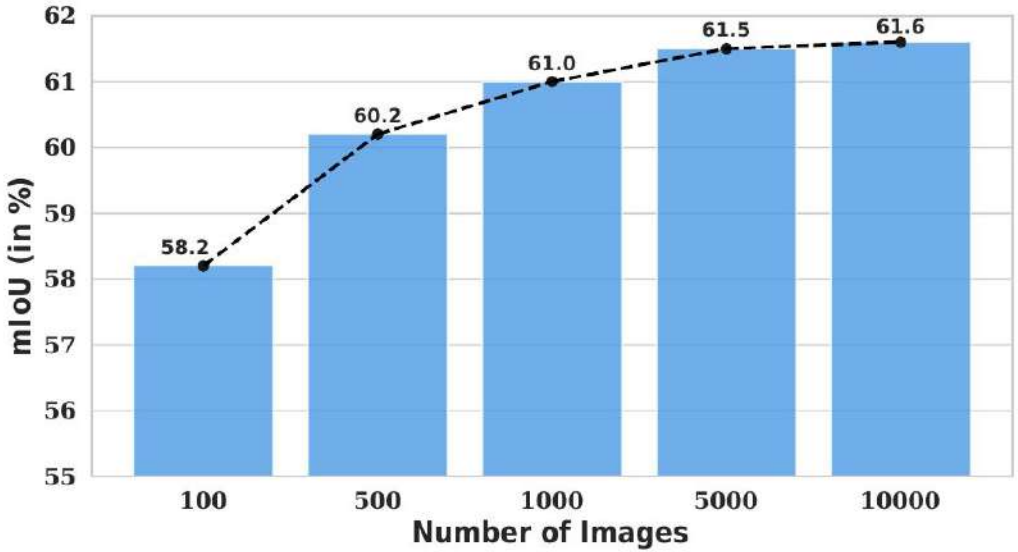
Ablation study

- Freezing the CLIP backbone helps to perform well in unseen domains

Backbone	Cityscapes	BDD100K	Mapillary	Avg.
Trainable	58.6	53.0	62.8	58.1
Frozen	**60.2**	**57.4**	**67.0**	**61.5**

Ablation study

- Increasing the size of the generated data helps to improve results until reaching a plateau



Conclusion

- Our method is one of the first to use FMs in DGSS, bridging a gap with the latest advancements in computer vision.

- CLOUDS integrates CLIP, diffusion models, LLMs, and SAM to enhance feature robustness, content and style diversity, and label refinement.

Future work

- Explore the **video** domain adaptation setting
- How can we better integrate **text modality** into existing architecture
- Explore the **robustness** of **open-vocabulary** methods to unseen domains

Thank you !