Adversarially Domain-adaptive Latent Diffusion for Unsupervised Semantic Segmentation

UNIVERSITY OF CAMBRIDGE

nextstud.io

KING'S College LONDON

Jongmin Yu [1,2] Zhongtian Sun [2,3] Chen Bene Chi [4] Jinhong Yan [1,5] Shan Luo [6]

[1]ProjectG.AI [2]University of Cambridge [3]University of Kent [4]Tsinghua University [5]Inje University [6]King's College London

Abstract

Semantic segmentation requires extensive pixel-level annotation, motivating unsupervised domain adaptation (UDA) to transfer knowledge from labelled source domains to unlabelled or weakly labelled target domains. However, learning a well-generalised representation that captures both domains remains very challenging, owing to probabilistic and geometric discrepancies between the virtual world and real-world imagery. This work introduces a semantic segmentation method based on latent diffusion models, termed Inter-Coder Connected Latent Diffusion (ICCLD), alongside an unsupervised domain adaptation approach. The model employs an inter-coder connection to enhance contextual understanding and preserve fine details, while adversarial learning aligns latent feature distributions across domains during the latent diffusion process. Experiments on GTA5, Synthia, and Cityscapes demonstrate that ICCLD outperforms state-of-the-art UDA methods, achieving mIoU scores of 74.4 (GTA5→Cityscapes) and 67.2 (Synthia→Cityscapes).

Introduction

Semantic segmentation, which involves assigning a semantic label to each pixel in an image, is crucial for applications such as autonomous driving. Despite significant advances in deep learning on standard benchmarks, many state-of-the-art (SOTA) models still struggle in real-world environments due to domain variations such as differences in lighting, geographic location, and weather conditions. Although increasing the diversity of training data through manual annotation can improve performance, the high cost, for example, approximately 90 minutes per image in Cityscapes [1], renders this approach impractical. Synthetic datasets such as Synthia [7] and GTA5 [6] offer a promising alternative. As these datasets are generated within controlled virtual environments, such as video games, pixel-level annotations can be automatically generated. However, the substantial visual disparity between simulated and real domains can significantly degrade the performance of models trained solely on synthetic data.

To address this issue, we propose a novel semantic segmentation framework integrating a diffusion model with adversarial learning. Specifically, we present Inter-Coder Connected Latent Diffusion (ICCLD), which incorporates long skip connections to bridge information between each encoder and decoder pair, effectively combining low- and high-level features with the latent representation produced by the diffusion process. Additionally, in the context of UDA, ICCLD employs a student and teacher learning framework, whereby the model iteratively refines its predictions on the target domain. Adversarial learning is integrated during the denoising phase, where a denoising UNet is trained to align feature distributions between the source and target domains by confusing a domain discriminator.

Summary of the key contributions

- **Latent diffusion-based semantic segmentation model (ICCLD):** We introduce ICCLD, which leverages long skip connections between encoder and decoder modules, combined with a conditioning mechanism, to enable the learning of more accurate semantic representations.
- **Improved domain adaptation via adversarial learning:** Our framework incorporates adversarial learning into the denoising process of the latent diffusion model, aligning latent feature distributions between the source and target domains.
- **Comprehensive empirical validation:** Extensive ablation studies and comparisons with existing methods demonstrate that ICCLD achieves superior performance on unsupervised domain adaptation benchmarks, with mIoU scores of 74.4 (GTA5→Cityscapes) and 67.2 (Synthia→Cityscapes).

References

[1] 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.

[2] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. *arXiv preprint arXiv:2111.14887*, 2021.

[3] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. *arXiv preprint arXiv:2204.13132*, 2022.

[4] Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8801–8811, 2021.

[5] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

[6] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[7] 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.

[8] Jongmin Yu, Hyeontaek Oh, and Jinhong Yang. Adversarial denoising diffusion model for unsupervised anomaly detection. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

Proposed Method

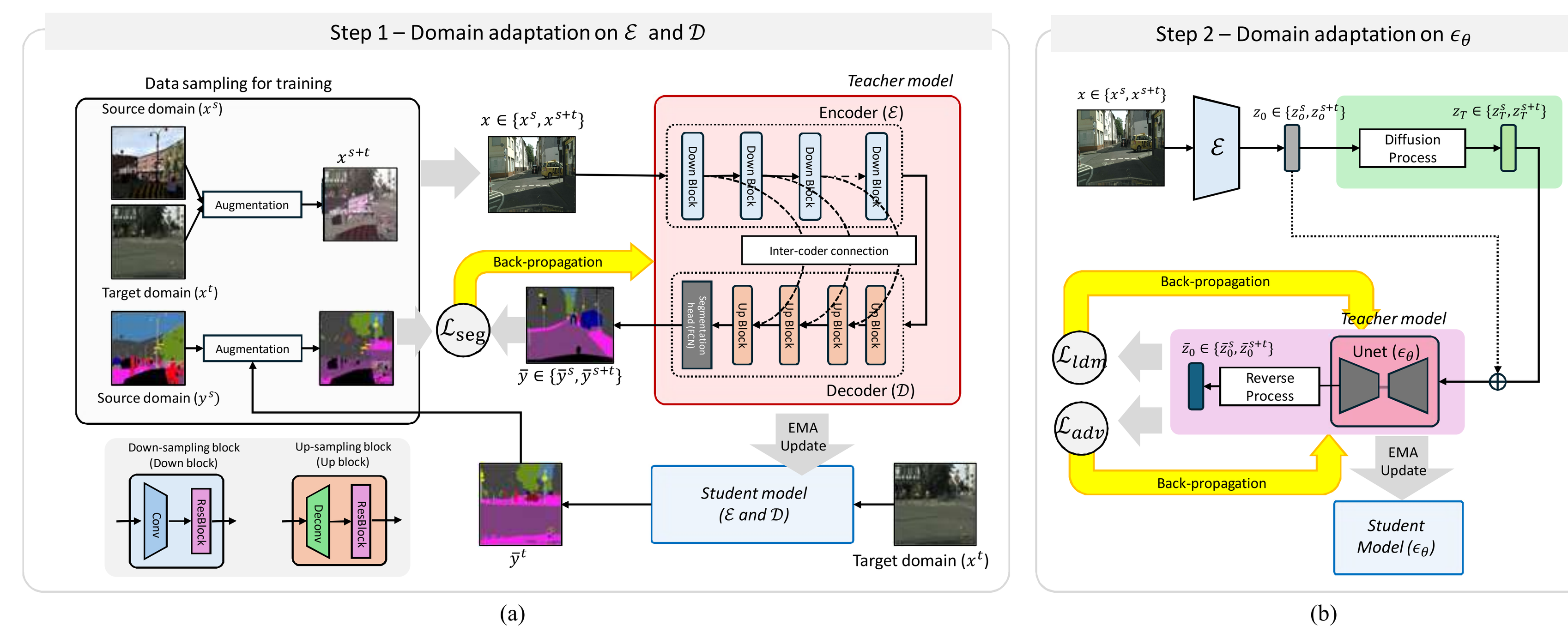

Figure 1. Illustration of the workflow for the two-step domain adaptation process using the proposed Inter-Coder Connected Latent Diffusion (ICCLD) framework.

Figure 1 illustrates our proposed approach for unsupervised domain adaptation using the Inter-Coder Connected Latent Diffusion (ICCLD) framework. The training process is conducted in two stages: (1) domain adaptation via segmentation, and (2) latent distribution alignment using a latent diffusion model. In the first stage, domain adaptation via segmentation is applied to derive a domain-adapted encoder $\mathcal{E}$ and decoder $\mathcal{D}$. These components are trained to minimise pixel-wise classification error using a segmentation-based approach. To facilitate this, we generate synthetic images and corresponding labels by combining annotated source domain data with unlabelled target domain data. In the second stage, latent distribution alignment is performed by training the denoising network ϵ_θ. The network is primarily optimised using the noise prediction loss defined by DDPM. To further reduce the domain discrepancy in the latent space, we introduce an additional loss term based on adversarial learning. The student model is updated via the Exponential Moving Average (EMA) of the teacher model's parameters.

Two-Step Unsupervised Domain Adaptation

Step 1-Domain adaptation on $\mathcal{E}$ and $\mathcal{D}$: After set a student model and a teacher model. Our framework takes as input a source image x^s and a target image x^t, sampled from a source dataset and a target dataset, respectively. A mixed image x^{s+t} is then generated using the ClassMix method [5], where source pixels from randomly selected classes are overlaid on the target image with pixel-level augmentations. This mixing process is equivalently applied to the corresponding labels. Since the target domain lacks annotations during training, pseudo labels y^{s+t} for x^{s+t} are generated by combining predicted labels $\bar{y}^t$ obtained from the student model and prepared target label y^s. The loss function using x^s, x^{s+t}, y^s, and y^{s+t} is formulated based on the cross-entropy function, which is defined as follows:

$$\mathcal{L}_{seg}(\bar{y}^*, y^*) = -\left[y^* \log(\bar{y}^*)\right], \tag{1}$$

where $\bar{y}^*$ can be defined by predicted segmentation results $\bar{y}^s$ and $\bar{y}^{s+t}$, and y^* can be defined by image labels y^s and y^{s+t}. Figure 1(a) provides a graphical explanation of the first step of DA for $\mathcal{E}$ and $\mathcal{D}$.

Step 2-Adversarial domain adaptation on ϵ_θ: The second phase is applied to improve domain adaptation performance by explicitly aligning latent feature distributions when we train the de-noising Unet ϵ^θ.

Since semantic segmentation is not just generating arbitrary segmentation masks. ICCLD should generate a suitable segmentation mask corresponding with the given image so that conditioning using the latent features is essential. As shown in Figure 1(b), the clear latent feature z_0, extracted by $\mathcal{E}$, is applied as a conditional factor during diffusion. The loss function for the t-step diffusion and de-noising processes is formulated as follows:

$$\mathcal{L}_{ldm}(t, z^*, z_0^*, \epsilon) = \left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} z^* + \sqrt{1-\bar{\alpha}_t}\epsilon, t|z_0^*)\right\|^2, \tag{2}$$

where z^* can be defined by the latent features for the source image z^s, mixed image z^{s+t}, and target image z^t. z_0^* indicates clear latent features that have no noise, and are extracted from source image z_0^s, target image z_0^{s+t}, and mixed image z_0^{s+t}. The scheduled variance defines α.

An adversarial loss is applied during every diffusion and denoising process to align the latent features from the source, mixed, and target domains. Adversarial learning is formulated to distinguish the predicted noise for each domain. We formulate the adversarial learning loss using KL-divergence. The adversarial loss is formulated as follows:

$$\mathcal{L}_{adv}(z^*, t) = \mathbb{E}\left[o^* \log\left(f_{dis}(\epsilon_\theta(\sqrt{\bar{\alpha}_t} z^* + \sqrt{1-\bar{\alpha}_t}\epsilon, t))\right)\right] + \mathbb{E}\left[D_{KL}\left(f_{dis}(\epsilon_\theta(\sqrt{\bar{\alpha}_t} z^* + \sqrt{1-\bar{\alpha}_t}\epsilon, t))|U(\frac{1}{3})\right)\right], \tag{3}$$

where f_{dis} defines a neural network that plays the role of a discriminator, and it outputs the results of the softmax function. z^* can be defined by the latent features for the source image z^s, mixed image z^{s+t}, and target image z^t. $o*$ defines pseudo-one-hot encoded vectors for indicating the domain of latent features used for predicting noise.

Experimental results

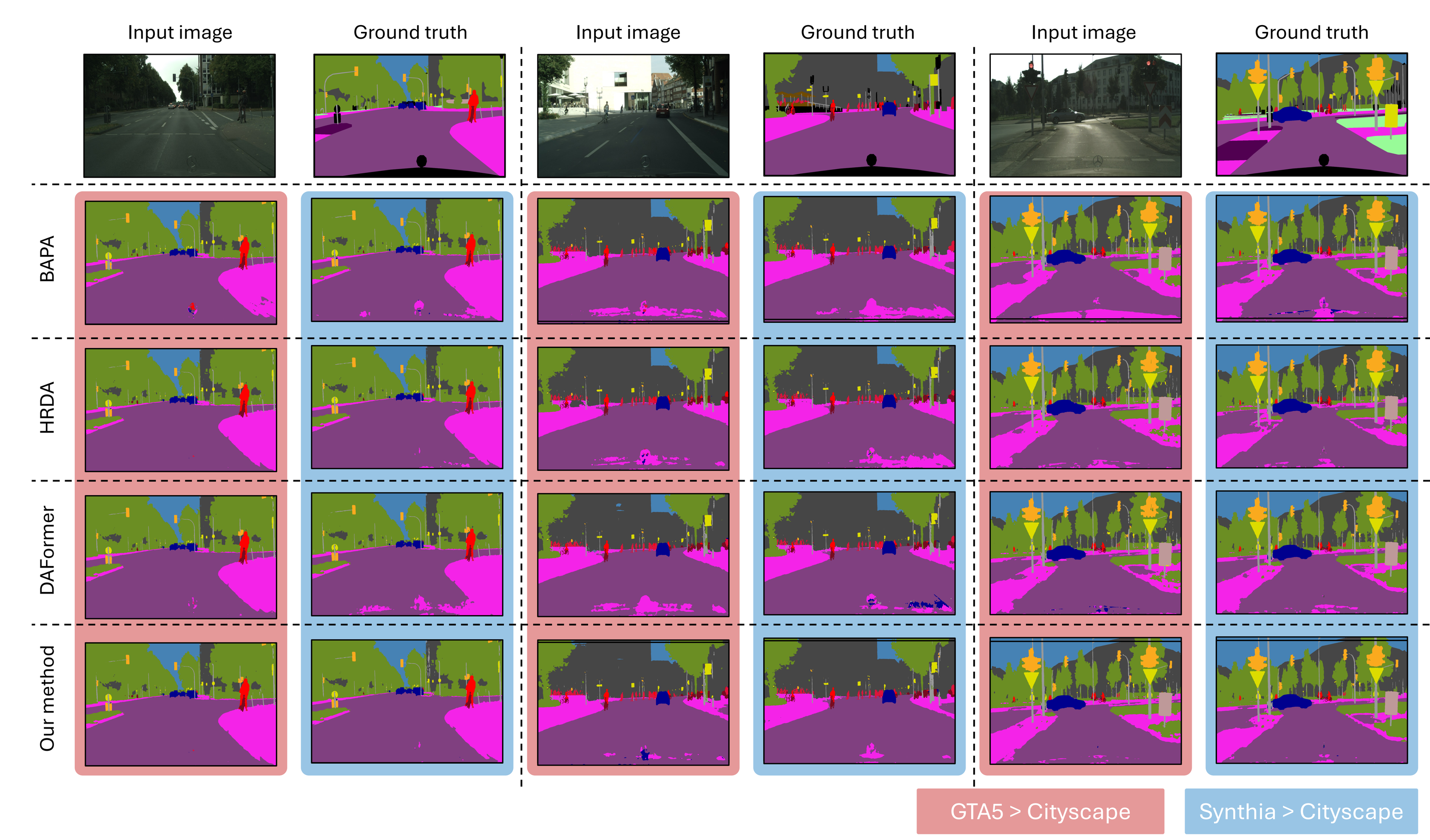

Figure 2. Qualitative comparison on the UDA performance using the proposed ICCLD with HRDA [3], DAFormer [2], and BAPA [4]. HRDA and DAFormer achieve the 2nd and 3rd ranked performances on mIoU (See Table 1).

Method	mIoU	Road	S.Walk	Build.	Wall	Fence	Pole	T. Light	T. Sign	Veget	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.Bike	Bike
GTA5 → Cityscapes																				
AdaptSeg	41.4	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5
CBST	45.9	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8
DACS	52.1	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
CorDA	56.6	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0
BAPA	57.4	94.4	61.0	88.0	26.8	39.9	38.3	46.1	55.3	87.8	46.1	89.4	68.8	40.0	90.2	60.4	59.0	0.0	45.1	54.2
ProDA	57.5	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
DAFormer	68.3	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
HRDA	73.8	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**
Our method	**74.4**	**97.6**	74.3	90.9	**62.3**	**52.3**	57.0	**64.9**	**72.5**	91.1	**51.3**	**94.5**	78.6	**53.2**	**94.5**	**84.9**	85.4	75.1	**65.4**	66.7
Synthia → Cityscapes																				
AdaptSeg	37.2	79.2	37.2	78.8	-	-	-	9.9	10.5	78.2	-	80.5	53.5	19.6	67.0	-	29.5	-	21.6	31.3
CBST	42.6	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	-	78.3	60.6	28.3	81.6	-	23.5	-	18.8	39.8
DACS	48.3	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	-	90.8	67.5	38.3	82.9	-	38.9	-	28.5	47.6
CorDA	55.0	93.3	61.6	85.3	19.6	5.1	37.8	36.6	42.8	84.9	-	90.4	69.7	41.8	85.6	-	38.4	-	32.6	53.9
BAPA	53.3	91.7	53.8	83.9	22.4	0.8	34.9	30.5	42.8	**86.8**	-	88.2	66.0	34.1	86.6	-	51.3	-	29.4	50.5
ProDA	55.5	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
DAFormer	60.9	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
HRDA	65.8	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
Our method	**67.2**	**91.3**	**48.5**	**89.9**	**49.5**	**8.5**	**58.6**	**67.4**	60.2	86.2	-	**93.4**	**79.8**	**54.2**	88.9	-	**66.7**	-	**65.5**	**65.7**

Table 1. Comparison with existing DA methods for UDA. The bolded figures represent the best performances.

Our method outperforms current SOTA techniques by a margin of +0.6 mIoU on the GTA5→Cityscapes benchmark and +1.4 mIoU on the Synthia→Cityscapes benchmark. Class-wise improvements are observed in 11 out of 19 classes for GTA5→Cityscapes—particularly in challenging categories such as Wall, Rider, and Fence—and in 12 out of 16 classes for Synthia→Cityscapes, as further quantified in Table 1.

Conclusion

In this paper, we introduce the Conditional and Inter-coder Connected Latent Diffusion (CICLD) model and adversarial learning for unsupervised domain-adaptive semantic segmentation. CICLD is a variation of latent diffusion models, and it has inter-coder connections which bridge information obtained by an encoder to a decoder. In training UNet for the latent diffusion, we apply adversarial learning to reduce domain gaps by aligning the probabilistic character of the predicted noise distribution. Experimental results on GTA5→Cityscapes and Synthia→Cityscapes benchmarks show that CICLD achieves mIoU scores of 74.4 and 67.2, respectively, outperforming current SOTA methods.

jy522@projectg.ai