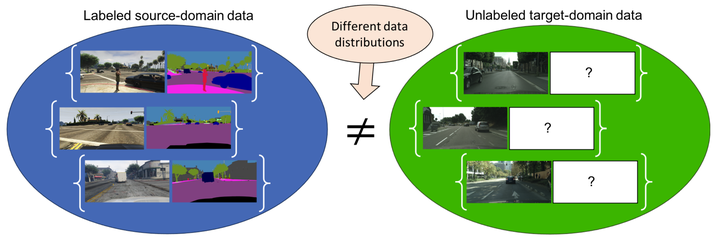
Beyond Deterministic Translation for Unsupervised Domain Adaptation

Eleni Chiou (UCL), Eleftheria Panagiotaki (UCL), Iasonas Kokkinos (UCL, Snap Inc.)

Problem Description

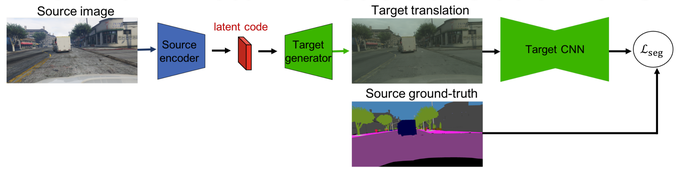
Target-domain data encountered at test time may have different distribution from source-domain data encountered at training time.



Problem: Models trained on source do not generalize well on the target domain.

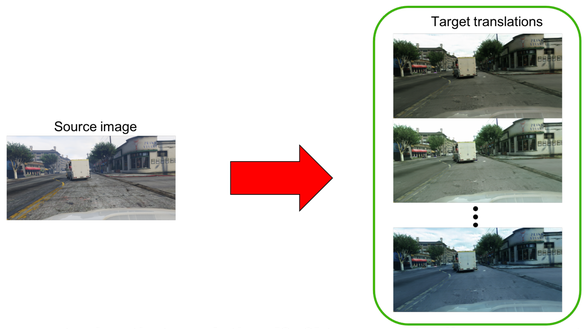
Potential Solution: Unsupervised domain adaptation (UDA) to adapt a source model to the target-domain dataset.

Prior work: Deterministic translation for UDA



- Translate a source-domain image to a single target-domain image relying on deterministic translation [1].
- Train a target network based on ground-truth originally available in the source domain.

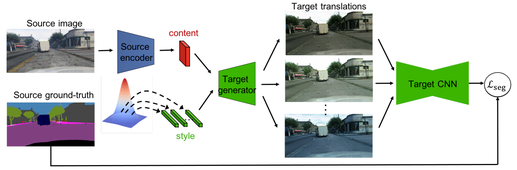
Our work: Stochastic translation for UDA



- Relies on **stochastic translation** [2, 3] to generate **diverse outputs** conditioned on the **same input**.

Proposed Method

Stochastic Translation and UDA

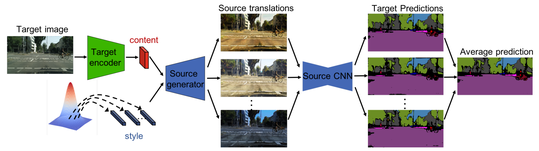


- Rely on a content-style disentanglement network [3].
- Associate a *source-domain* image, $x \in \mathcal{S}$, with a *distribution* of image translations to the *target domain*:

$$\mathbf{T}[x, v] \doteq G_t(C_s(x), v), v \sim \mathcal{N}(0, \mathbf{I}), x \in \mathcal{S},$$

where we encode the content of the source image through $C_s(x)$ and then pass it to the target-domain generator G_t that is driven by the random style code v.

Stochastic Translation and Pseudo-labelling

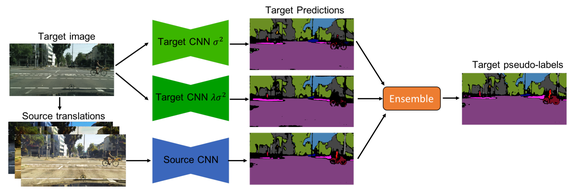


- Consider a complementary source CNN, F_s, that operates in the source domain.
- Obtain Monte Carlo estimate of pseudo-labels by exploiting the whole distribution of image translations *from the target to the source domain*:

$$\hat{y}(x) = E_v[F_s(\mathbf{I}[x, v])], x \in \mathcal{T}, v \sim \mathcal{N}(0, I) \approx \frac{1}{K}\sum_{k=1}^{K} F_s(I[x, v_k])$$

where $\mathbf{I}$ is the inverse transform from the target domain, $\mathcal{T}$, to the source and v_k is independently sampled from the normal distribution.

Ensemble of a triplet of networks



- Train two target networks, one with the variance left intact and the other with the variance scaled by a constant.
- Average their predictions with those of the source-domain network described previously.

Results

Ablation studies on GTA-to-Cityscapes

- Stochastic vs deterministic translation

Method	Output space	Pixel space	mIoU
ADVENT [4]	✓		43.8
ADVENT *	✓		42.9
ADVENT *+ CycleGAN*	✓	✓	45.1
Ours	✓	✓	46.2
Ours w/ L_{sem}	✓	✓	**46.6**

Train ADVENT[4] using synthetic images obtained from deterministic translation (CycleGAN) and stochastic translation (Ours). Stochastic translation improves performance. $*$ denotes our retrained models.

- Robust pseudo-labelling through network ensemble

F_s, K=1	F_s, K=5	F_s, K=10	F_t, $\sigma^2=1$	F_t, $\sigma^2=10$	mIoU
✓					43.3
	✓				44.0
		✓			44.4
			✓		46.6
				✓	46.1
		✓	✓		47.7
		✓		✓	47.6
			✓	✓	47.7
		✓	✓	✓	**48.2**

Rows 1-3: performance of the source network F_s when averaging the predictions of multiple translations K, of a target image. Rows 4-5: performance of the target networks F_t, trained with different degrees of stochasticity in the translation. Row 6-8: performance when averaging the predictions.

Benchmark results

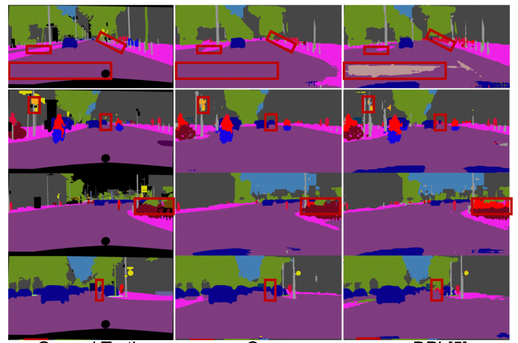
Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
ResNet101 backbone																				
AdvEnt[30]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
BDL [16]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
LTIR [14]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
FDA-MBT [36]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
PCEDA [37]	91.0	49.2	85.6	37.2	29.7	33.7	38.1	39.2	85.4	35.4	85.1	61.1	32.8	84.1	**45.6**	46.9	0.0	34.2	44.5	50.5
TPLD [26]	**94.2**	**60.5**	82.8	36.6	16.6	39.3	29.0	25.5	85.6	44.9	84.4	60.6	27.4	84.1	37.0	47.0	**31.2**	36.1	**50.3**	51.2
Wang et al. [32]	90.5	38.7	**86.5**	41.1	32.9	**40.5**	48.2	42.1	86.5	36.8	84.2	**64.5**	**38.1**	87.2	34.8	50.4	0.2	41.8	54.6	52.6
PixMatch [21]	91.6	51.2	84.7	37.3	29.1	24.6	31.3	37.2	86.5	44.3	85.3	62.8	22.6	87.6	38.9	52.3	0.65	37.2	50.0	50.3
DPL-Dual (Ensemble) [5]	92.8	54.4	86.2	41.6	32.7	36.4	**49.0**	34.0	85.8	41.3	**86.0**	63.2	34.2	87.2	39.3	44.5	18.7	**42.6**	43.1	53.3
SUDA [38]	91.1	52.3	82.9	30.1	25.7	38.0	44.9	38.2	83.9	39.1	79.2	58.4	26.4	84.5	37.7	45.6	10.1	23.1	36.0	48.8
CaCo [11]	91.9	54.3	82.7	31.7	25.0	38.1	46.7	39.2	82.6	39.7	76.2	63.5	23.6	85.1	38.6	47.8	10.3	23.4	35.1	49.2
Ours	93.3	56.5	85.9	41.0	33.1	34.8	43.8	**43.8**	86.6	46.5	82.5	61.1	30.4	87.0	39.7	50.7	8.8	34.9	46.8	53.0
Ours (Ensemble)	93.4	55.8	86.4	**44.4**	**36.1**	34.6	45.0	39.8	**86.9**	**48.0**	84.4	61.7	30.9	**87.7**	44.9	**55.9**	11.1	38.4	45.4	**54.3**

Quantitative comparison on GTA5→Cityscapes. Per-class IoU and mean IoU (mIoU) obtained using VGG and ResNet101 backbones.

Method	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	sky	person	rider	car	bus	motorcycle	bicycle	mIoU	mIoU*
ResNet101 backbone																		
AdvEnt[30]	85.6	42.2	79.7	-	-	-	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	-	48.0
LTIR [14]	92.6	53.2	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3
BDL [16]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
FDA-MBT [36]	79.3	35.0	73.2	-	-	-	19.9	24.0	61.7	82.6	61.4	**31.1**	83.9	40.8	**38.4**	51.1	-	52.5
PCEDA [37]	85.9	44.6	80.8	-	-	-	24.8	23.1	79.5	83.1	57.2	29.3	73.5	34.8	32.4	48.2	-	53.6
TPLD [26]	80.9	44.3	82.2	19.9	0.3	**40.6**	20.5	**30.1**	77.2	80.9	60.6	25.5	84.8	41.1	24.7	43.7	47.3	53.5
Wang et al. [32]	79.4	34.6	**83.5**	**19.3**	**2.8**	35.3	**32.1**	26.9	78.8	79.6	**66.6**	30.3	86.1	36.6	19.5	**56.9**	48.0	54.6
PixMatch [21]	**92.5**	**54.6**	79.8	4.7	0.08	24.1	22.8	17.8	79.4	76.5	60.8	24.7	85.7	33.5	26.4	54.4	46.1	54.5
DPL-Dual (Ensemble) [5]	87.5	45.7	82.8	13.3	0.6	33.2	22.0	20.1	83.1	86.0	56.6	21.9	83.1	40.3	29.8	45.7	47.0	54.2
SUDA [38]	83.4	36.0	71.3	8.7	0.1	26.0	18.2	26.7	72.4	80.2	58.4	30.8	80.6	38.7	36.1	46.1	44.6	52.2
CaCo [11]	87.4	48.9	79.6	8.8	0.2	30.1	17.4	28.3	79.9	81.2	56.3	24.2	78.6	39.2	28.1	48.3	46.0	53.6
Ours	85.8	41.7	82.4	7.6	1.9	33.2	26.5	18.4	83.3	86.5	62.0	29.7	83.9	52.1	34.6	51.4	48.8	56.8
Ours (Ensemble)	87.2	44.1	82.1	6.5	1.4	33.1	24.7	17.9	**83.4**	**86.6**	62.4	30.4	**86.1**	**58.5**	36.8	52.8	**49.6**	**57.9**

Quantitative comparison on GTA5→Cityscapes. Per-class IoU and mean IoU (mIoU) obtained using VGG and ResNet101 backbones.

Qualitative Results



1. Zhu et al., Unpaired image-to-image translation using cycle-consistent adversarial networks, ICCV 2017
2. Huang et al., Multimodal Unsupervised Image-to-Image Translation, ECCV 2018
3. Lee et al., Diverse Image-to-Image Translation via Disentangled Representations, ECCV 2018
4. Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation, CVPR 2019
5 Cheng et al., Dual Path Learning for Domain Adaptation of Semantic Segmentation, ICCV 2021

Code: https://github.com/elchiou/Beyond-deterministic-translation-for-UDA