Category-Adaptive Domain Adaptation for Semantic Segmentation

Authors: Zhiming Wang, Yantian Luo, Danlan Huang, Ning Ge, Jianhua Lu

Department of Electronic Engineering, Tsinghua University, Beijing 100084, P.R. China
Beijing National Research Center for Information Science and Technology

Outline

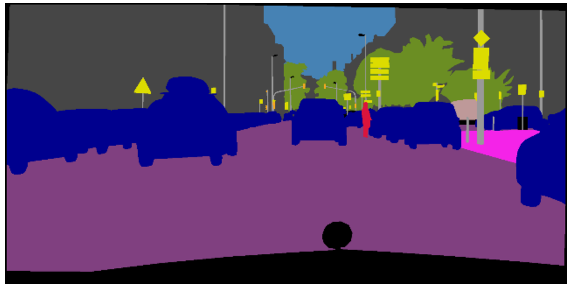
- Introduction
  - Backgrounds
  - Problem statement
  - Baseline & Proposed method
- Experiments
  - Datasets
  - Training settings & Loss functions
  - Quantitative performance
  - Qualitative performance
  - Ablation Study
- Conclusions
  - Takeaways
  - Future work

Introduction / Backgrounds

Semantic Segmentation

- Applications:
  - Auto driving
  - Scene understanding
  - Medical diagnosis
- Categories
  - Supervised manner
    - Advantages: excellent performance and model is easy to train
    - Disadvantages:

Image



Semantic map

No annotations

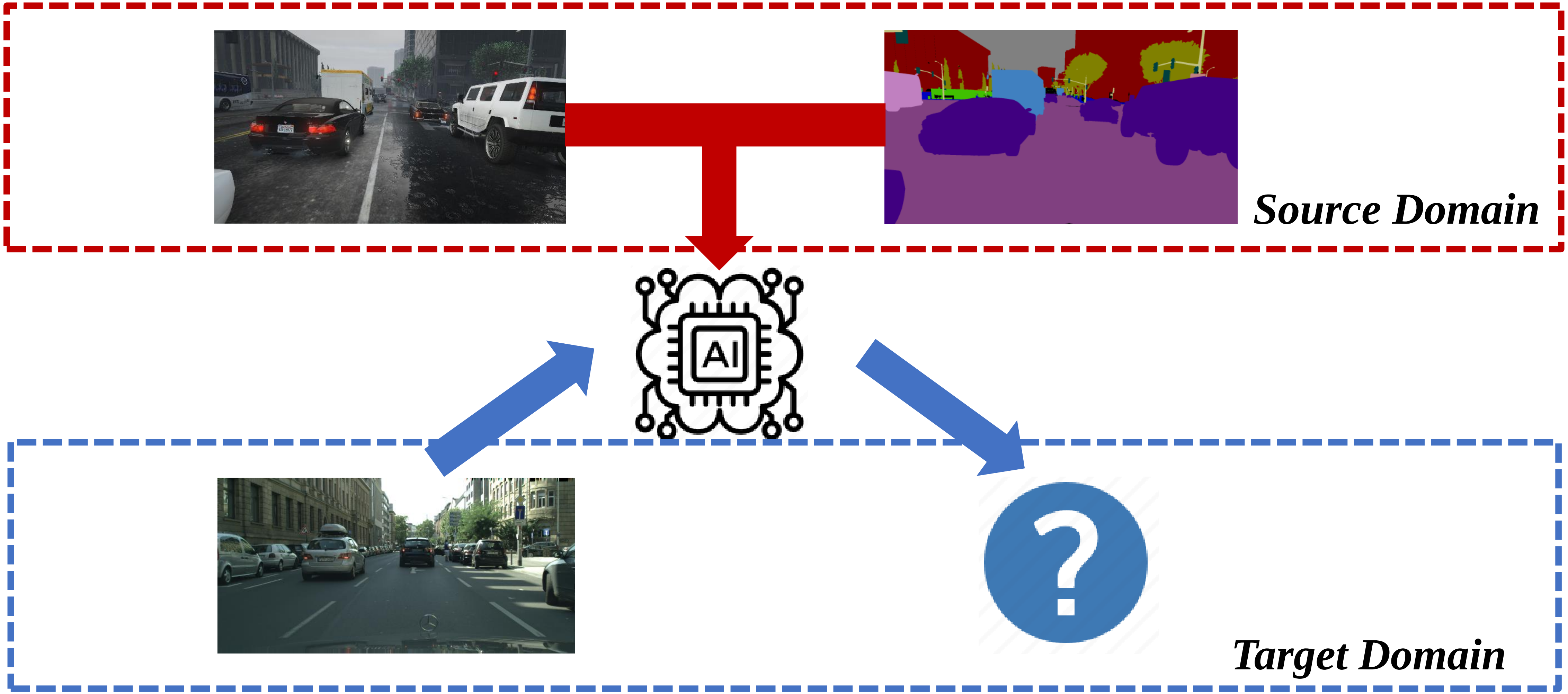
Tedious

Time consuming

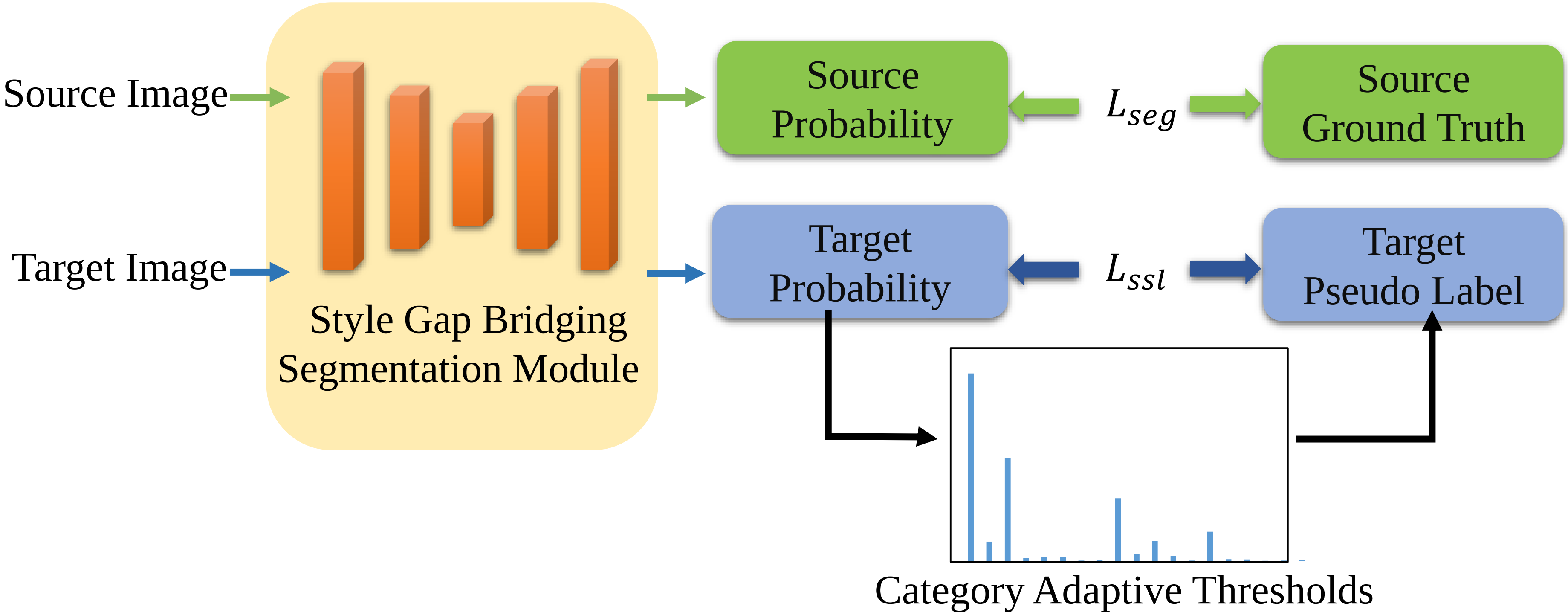
  - Domain adaptation based semantic segmentations

Introduction / Problem Statement

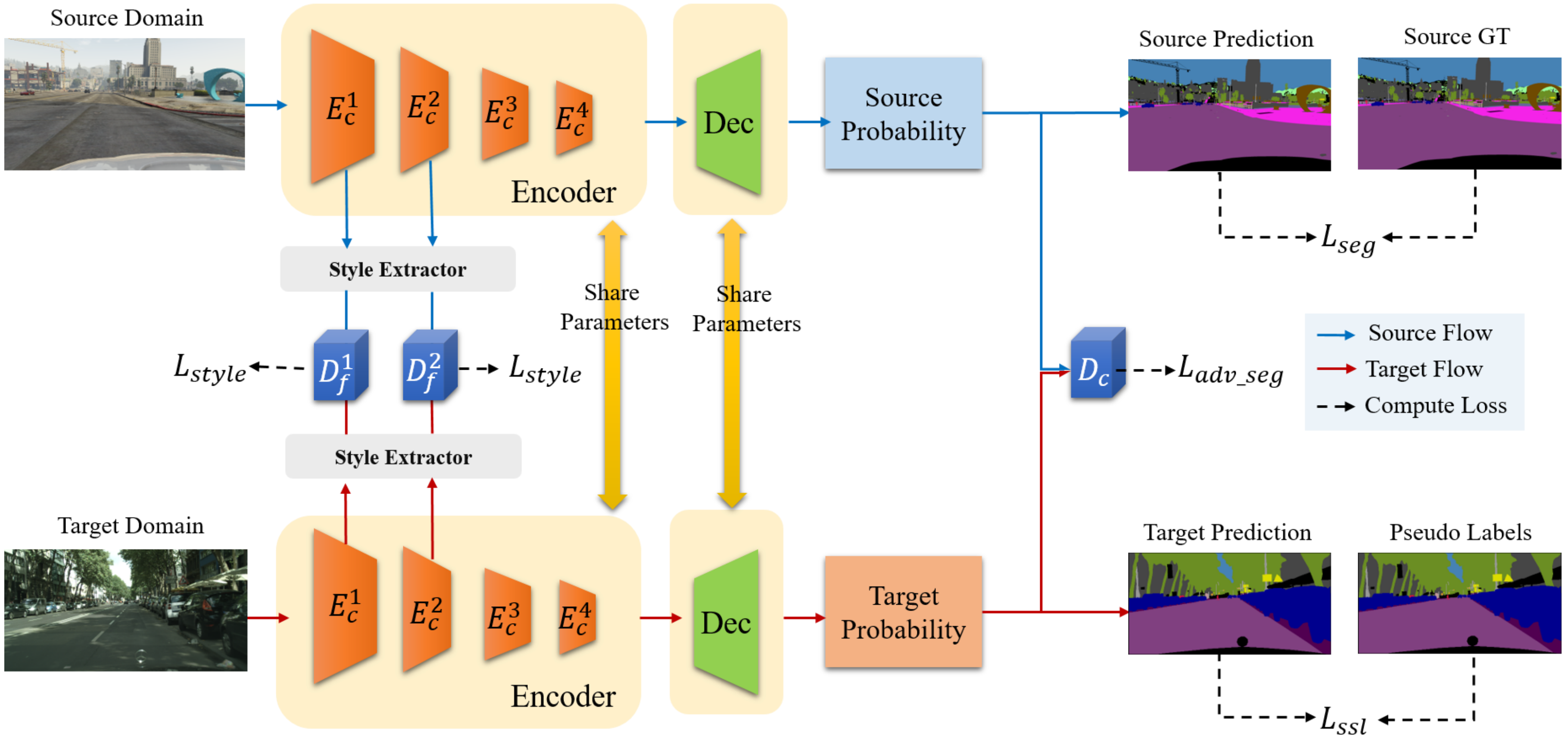
- Domain adaptation based semantic segmentations



- Pipeline of the cross-domain adaptation for semantic segmentation:



Introduction / Proposed Method



D_c: segmentation discriminator D_f^1, D_f^2: style discriminators

Tsinghua University

Style Gap Bridging Mechanism:

- Previous work: MSE of the channel-wise statistics of extracted features
- Our work:
  - Feature-level: adversarial training of the channel-wise mean of extracted features
  - Output-level: adversarial training of the output probability maps
- Motivation:
  - MSE requires the features meet with Gaussian distribution assumption
  - Adversarial training is proved to narrow the distribution distance of data

Introduction / Proposed Method

- Previous pseudo labeling:
  - set a fixed threshold for all categories (like BDL)
  - leverage category-wise ratio priors (like ADVENT, CBST)

- category-adaptive threshold mechanism for pseudo labeling:
  - Given $P_t \in \mathbb{R}^{H_t \times W_t \times C}$, the category centroid is defined as follows:

$$f^c = \frac{1}{|\mathcal{P}^c|} \sum_{|\mathcal{X}_t|} \sum_{h=1}^{H_t} \sum_{w=1}^{W_t} \mathbb{1}_{[c=\arg\max_{c'} P_t^{hwc'}]} P_t^{hw}$$

  - an indicator variable is defined as follows:

$$m_t^{hwc} = \mathbb{1}_{[H(P_t^{hw}) < H(f^c) - \Delta,\, c=\arg\max_{c'} P_t^{hwc'}]}$$

    - $H(\cdot)$ denotes the entropy function
    - Δ is a manually fixed hyper-parameter to control the threshold for each category.

Experiments / Loss Functions

Two training phases: Domain adaptation training and SSL

- Domain adaptation training phase
  - Segmentation Loss:

$$\mathcal{L}_{seg} = -\frac{1}{H_s W_s} \sum_{h=1}^{H_s} \sum_{w=1}^{W_s} \sum_{c=1}^{C} y_s^{hwc} \log \hat{y}_s^{hwc}$$

  - Output-based Domain Adaptation Loss:

$$\mathcal{L}_{adv\_seg} = -\min_{E_c, Dec} \max_{D_c} \mathbb{E}_{I_t \sim T} \log \left[D_c \left(\mathrm{Dec} \left(E_c \left(I_t \right) \right) \right) \right] + \mathbb{E}_{I_s \sim S} \log \left[1 - D_c \left(\mathrm{Dec} \left(E_c \left(I_s \right) \right) \right) \right]$$

  - Style Loss:

$$\mathcal{L}_{\text{style}} = -\sum_{m=1}^{M} \min_{E_c^m} \max_{D_f^m} \left\{ \mathbb{E}_{I_t \sim T} \log \left[D_f^m \left(S_{tm} \right) \right] + \mathbb{E}_{I_s \sim S} \sum_{m=1}^{M} \log \left[1 - D_f^m \left(S_{sm} \right) \right] \right\}$$

- Final loss on the domain adaptation training phase:

$$\mathcal{L} = \lambda_{seg} \mathcal{L}_{seg} + \lambda_{adv\_seg} \mathcal{L}_{adv\_seg} + \lambda_{style} \mathcal{L}_{style}$$

Two training phases: Domain adaptation training and SSL

- SSL phase
  - Self-supervised Loss:

$$\mathcal{L}_{ssl} = -\frac{1}{H_t W_t}\sum_{h=1}^{H_t}\sum_{w=1}^{W_t}\sum_{c=1}^{C} m_t^{hwc}\hat{y}_t^{hwc}\log P_t^{hwc}$$

- Final Loss during the SSL phase:

$$\mathcal{L} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{adv\_seg}\mathcal{L}_{adv\_seg} + \lambda_{style}\mathcal{L}_{style} + \mathcal{L}_{ssl}$$

Experiments / Datasets

- Source Domain Dataset: GTA5
  - 24966 synthetic images collected from the game engine
  - 19-category pixel-accurate annotations （compatible with Cityscapes）
- Target Domain Dataset: Cityscapes
  - collected from streetscapes in 50 different Germany cities
  - 2975 training images
  - 500 validation images (as the testing set)
  - 1525 testing images (abandoned for the lack of annotations)

Experiments / Training Settings

- Encoder architecture: DeepLab V2
- Segmentation and style discriminators' architecture: PatchGAN
- Hyper-parameters: $\lambda_{seg} = 1, \lambda_{adv\_seg} = \lambda_{style} = 1 \times 10^{-3}$

Module	Optimizer	Original learning rate	Leanring rate update
Encoder	SGD with momentum=0.9	2.5×10^{-4}	poly decay policy: maxstep=250,000 Power=0.9
Decoder		2.5×10^{-3}	
Discriminator	Adam with $\beta =$ (0.9,0.99)	1×10^{-4}	exponential decay policy: decay rate:0.1 decay steps: 50,000

Experiments / Quantitative Performance

Table 1: Comparison among different methods for "GTA5 to Cityscapes"

	GTA5 → Cityscapes																			
Method	road	sidewalk	building	wall	fence	pole	t-light	t-sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mIoU
CBST[4]	89.6	**58.9**	78.5	33.0	22.3	**41.4**	**48.2**	39.2	83.6	24.3	65.4	49.3	20.2	83.3	39.0	48.6	**12.5**	20.3	35.3	47.0
Cycada [19]	86.7	35.6	80.1	19.8	17.5	38.0	39.9	**41.5**	82.7	27.9	73.6	**64.9**	19	65.0	12.0	28.6	4.5	31.1	42.0	42.7
ADVENT [6]	87.6	21.4	82.0	34.8	26.2	28.5	35.6	23.0	84.5	35.1	76.2	58.6	30.7	84.8	34.2	43.4	0.4	28.4	35.2	44.8
DCAN [20]	85.0	30.8	81.3	25.8	21.2	22.2	25.4	26.6	83.4	36.7	76.2	58.9	24.9	80.7	29.5	42.9	2.5	26.9	11.6	41.7
CLAN [21]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	**31.9**	31.4	43.2
BDL [5]	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
Ours	**91.7**	51.1	**85.0**	**38.7**	26.7	32.1	38.1	34.6	84.3	38.6	**84.9**	60.7	**32.8**	**85.2**	**41.9**	**49.8**	2.8	28.5	**45.0**	**50.2**

- Compared with BDL, our method has a gain of 1.7 on overall mIoU.
- Compared with CBST, our model brings +3.2% mIoU improvement.
- Compared with ADVENT, our model brings +5.4% mIoU improvement.

Experiments / Ablation Study

Table 2: Ablation study on SSL and style constraints.

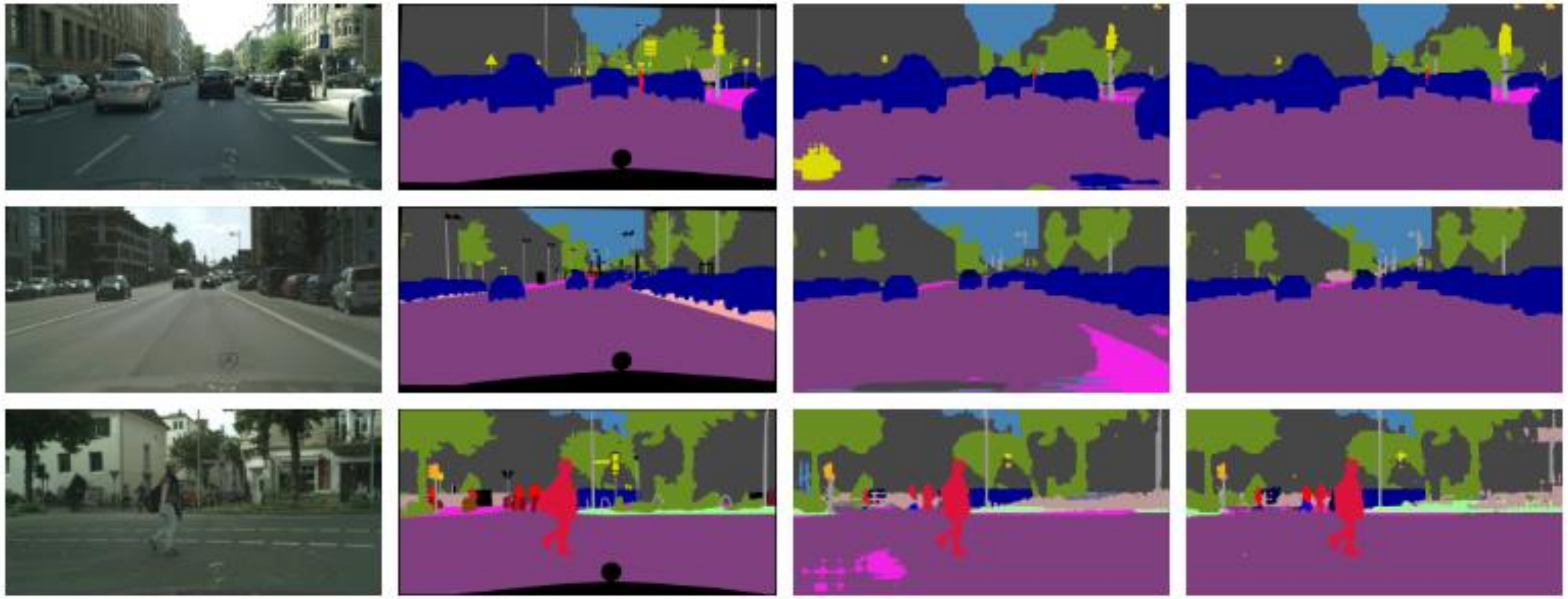
GTA5 $\rightarrow$ Cityscapes	
model	mIoU
original	44.6
original + adv	45.5
original + adv + SSL once	48.5
original + adv + SSL twice	50.2

Experiments / Ablation Study

Table 3: Comparison on style gap bridging mechanisms

style gap bridging mechanism	style modeling	mIoU
MSE	Gram matrix	44.7
	mean & std	45.1
adversarial learning	mean (Ours)	45.5

Experiments / Qualitative Performance



(a) Image (b) GT (c) BDL (d) Ours

Conclusions

- Takeaways:
  - propose a style gap bridging mechanism based on adversarial learning
  - propose a category-adaptive threshold mechanism to choose pseudo labels for SSL
- Future work:
  - an elaborate network architecture is worth exploring
  - an efficiency pseudo labeling mechanism is appealing
  - the statistic modeling of “style information” needs further research

References

[1] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in Proc. ICCV, 2011, pp. 2018–2025.
[2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Aneural algorithm of artistic style,"arXiv preprintarXiv:1508.06576, 2015.
[3] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proc. ICCV, 2017, pp. 1501–1510.
[4] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in Proc. ECCV, 2018,pp. 289–305
[5] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in Proc. CVPR, 2019, pp. 6936–6945.
[6] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. P´erez,"Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in Proc. CVPR,2019, pp. 2517–2526.
[7] Hou and L. Zheng, "Source free domain adaptation with image translation,"arXiv preprintarXiv:2008.07514, 2020.
[8] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, "Correntropy induced l2 graph for robust subspace clustering," in Proc. ICCV, 2013, pp. 1801–1808.
[9] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for se-mantic segmentation," in Proc. NIPS, 2019, pp. 435–445.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,"Generative adversarial nets," in Proc. NIPS 2014, pp.2672–2680.
[11] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in Proc. ECCV, 2016, pp. 102–118.
[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. En-zweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proc. CVPR, 2016, pp. 3213–3223.
[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy,and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, "IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp.834–848, 2017
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," inProc. CVPR, 2016,pp. 770–778.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Ima-genet classification with deep convolutional neural networks,"Communications of the ACM, vol. 60, no. 6, pp.84–90, 2017.
[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial net-works," in Proc. VPR, 2017, pp. 1125–1134.

清華大學
Tsinghua University

Thank you！

wang-zm18@mails.tsinghua.edu.cn