CATEGORY-ADAPTIVE DOMAIN ADAPTATION FOR SEMANTIC SEGMENTATION

Zhiming Wang, Yantian Luo, Danlan Huang, Ning Ge, Jianhua Lu
Department of Electronic Engineering, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, Beijing, China

ABSTRACT
Unsupervised domain adaptation (UDA) becomes more and more popular in tackling real-world problems without ground truth of the target domain. Though tedious annotation work is not required, UDA unavoidably faces two problems: 1) how to narrow the domain discrepancy to boost the transferring performance; 2) how to improve pseudo annotation producing mechanism for self-supervised learning (SSL). In this paper, we focus on UDA for semantic segmentation task. Firstly, we introduce adversarial learning into style gap bridging mechanism to keep the style information from two domains in the similar space. Secondly, to keep the balance of pseudo labels on each category, we propose a category-adaptive threshold mechanism to choose category-wise pseudo labels for SSL. The experiments are conducted using GTA5 as the source domain, Cityscapes as the target domain. The results show that our model outperforms the state-of-the-arts with a noticeable gain on cross-domain adaptation tasks.

Index Terms— unsupervised domain adaptation, semantic segmentation, self-supervised learning

1. INTRODUCTION
As a significant task in computer vision, semantic segmentation aims at producing pixel-wise labels for images, and has been widely applied to many different scenes such as autonomous driving and scene understanding. However, semantic segmentation usually yields unsatisfying performance without enough labeled training samples. What’s more, it’s very difficult to apply supervised semantic segmentation to the emerging diverse applications. Thus, the pixel-wise annotations are time-consuming and expensive. Therefore, supervised learning based methods are unable to meet the requirement of current image segmentation tasks.

Domain adaptation (DA) offers a solution for semantic segmentation without huge amount of labeled training samples. It aims to apply a model pretrained on the source dataset to generalize on the target dataset. However, there usually exists huge gaps among datasets, which can be categorized into two folds: content-based gap and style-based gap. Content-based gap is caused by inter-dataset amount and frequency discrepancy of categories, which can be alleviated by choosing datasets with similar scenes so that it is often neglected for convenience. The style-based gap refers to the difference of illumination, things’ texture and so on. However, modelling the style information is still an open academic problem. It has been illustrated that the shallow layers of CNN extract low-level features (e.g., edges) while the deep layers extract high-level features (e.g., objects) [1]. Based on the fact that different convolutional kernels are computed independently, most literatures regard channel-wise statistics of extracted features as the style information, such as correlation-based Gram matrix [2], means and standard deviations (evaluated by AdaIN [3]). Without loss of generality, in this paper, we adopt the means to model the style information by global average pooling process. However, narrowing the content-based gap is still full of challenges.

Moreover, great advances have been achieved on domain adaptation with SSL, whose key is pseudo labeling mechanism. It solves the problem of lacking available annotations on the target domain. CBST [4] introduces the amount of each category as one optimization term so as to balance the probability of pseudo labels of each category. However, each iteration of SSL requires the ordering operation, which is time-consuming. BDL [5] directly sets a fixed confidence threshold for all categories, and the pseudo labels are obtained when corresponding confidence scores are above such a threshold. However, the fixed threshold mechanism suffers from varying numbers of pseudo labels for different categories, which unavoidably hurts the final segmentation performance. ADVENT [6] introduces the category-wise ratio priors on source domain to guide the pseudo label selection. Nevertheless, it still remains challenging to avoid choosing pseudo labels biased towards easy categories.

In this paper, to address the above issues, we propose a category-adaptive domain adaptation approach for seman-
Fig. 2: The framework of our proposed model. Noting that content-based gap is not taken into account, here we take two datasets with regard to urban streetscape for example. The blue flow shows the process of the source domain images, while the red flow shows the process of the target domain images. The encoder module and decoder module are shared for two domain images, where the style extractor is achieved by global average pooling process. The model is firstly trained on the source domain in a supervised manner. Then it is trained on the target domain with SSL, where the pseudo labels for each category are filtered by the corresponding category-adaptive threshold.

The main contributions of this paper are summarized as follows:

1. We propose a style gap bridging mechanism based on adversarial learning, which narrows the style-based gap to help alleviate the domain discrepancy.
2. We propose a category-adaptive threshold mechanism for pseudo labeling to help SSL on the target domain images.
3. We conduct a series of experiments on cross-domain segmentation task and verify the effectiveness and superiority of our method.

2. PROPOSED METHOD

The framework of our proposed model is shown in Fig. 2. Firstly, the model is trained on both domain data, where the style information between both domains is aligned as close as possible. It is worth noticing that data of the target domain are lack of pixel-wise domain annotations. Then pseudo labels are chosen based on the prediction of pretrained model on target domain. At last, SSL is conducted on the target domain by virtue of chosen pseudo labels.

2.1. Style Gap Bridging Mechanism

The core of our encoder is to keep content information, meanwhile, decrease style information as much as possible, since the semantic performance heavily depends on content information. Therefore, it is reasonable to narrow the gaps of style information between source domain images and target domain images. In this paper, without loss of generality, we leverage global average pooling as the style extractor in Fig. 2, since channel-wise statistics are demonstrated related to style information [3]. Previous works [7,8] usually apply MSE as style constraints, however, MSE performs worse on data with high dimensions and is limited by the linearity and Gaussianity assumptions [9]. By contrast, adversarial learning is theoretically proved to narrow the gap between two high-dimensional distributions. In practice, with the help of style discriminators (i.e., $D_1^s$ and $D_2^t$), we apply adversarial loss on style information $S_{an}$ extracted from 2 front sub-encoder modules (i.e., $E_1^s$ and $E_2^t$ in Fig.2, where $* = s/t$ denotes the source domain / target domain, $n = \{1, 2\}$.

2.2. Pseudo labeling for target domain

Here we propose a category-adaptive threshold method for SSL. The idea is based on the hypothesis that the pretrained model’s performances on different categories are different because of the uneven prior distributions of different categories. For example, the category "road" accounts a lot
while the category “train” is just the reverse. Therefore, the confidence threshold should vary among different categories. Based on the clustering method of \cite{10} where the threshold is defined by the Euclidean distance between target features and category centroids, we consider that each intra-category feature makes different contributions to the category centroids because the prediction confidence varies. Consequently, based on the given model’s output on the target domain $P_t \in \mathbb{R}^{H_t \times W_t \times C}$, we firstly define a confidence-weighted target domain-based category centroid $f^l \in \mathbb{R}^C$:

$$f^l = \frac{1}{|P^l|} \sum_{h=1}^{H_t} \sum_{w=1}^{W_t} \sum_{c=1}^{C} \hat{y}_{hwc}^l P_t^{hwc},$$

(1)

where $P^l$ denotes the collection of prediction confidence of all pixels decided as $l$-th category, $|P^l|$ denotes the cardinality of $P^l$. $\hat{y}_{hwc}^l = 1_{[c=\text{arg max}_c p_y^l(hwc)]}$, and 1 is the binary indicator function.

Given $f^l$ in each category, our threshold is based on the entropy distance. The entropy of prediction vector at the $h$th row and with column $P_t^{hwc} \in \mathbb{R}^C$ is:

$$E(P_t^{hwc}) = -\sum_{c=1}^{C} P_t^{hwc} \log P_t^{hwc}. \quad (2)$$

The entropy of category centroid $f^l$, namely $E(f^l)$ is similar with Equation (2). Intuitively, $E(P_t^{hwc})$ decreases as the max confidence in $P_t^{hwc}$ increases, consequently we choose entropy-based threshold. Here we defined an indicator variable $m_t^{hwc}$ to decide whether the prediction on current position is chosen as available pseudo labels:

$$m_t^{hwc} = 1_{[E(P_t^{hwc})<E(P_f^l)-\Delta]}, \quad (3)$$

where $\Delta$ is a manually fixed hyperparameter to control the threshold for each category. When $\Delta$ increases, the number of available pseudo labels decreases while the model will have higher prediction confidence and vice versa.

### 2.3. Loss Functions

As mentioned above, the training process includes two phases: domain adaptation training and SSL. Domain adaptation training process utilizes the following three losses:

#### Segmentation Loss. Here cross entropy function is applied to penalize the error between prediction $\hat{y}_s \in \mathbb{R}^{H_s \times W_s \times C}$ and one-hot ground truth $y_s \in \mathbb{R}^{H_s \times W_s \times C}$:

$$L_{seg} = -\frac{1}{H_s \times W_s} \sum_{h=1}^{H_s} \sum_{w=1}^{W_s} \sum_{c=1}^{C} y_{s}^{hwc} \log \hat{y}_{s}^{hwc}. \quad (4)$$

#### Output-based Domain Adaptation Loss. Consistent with BDL \cite{5}, we also leverage the original GAN loss introduced by Goodfellow \cite{11} as $L_{adv,seg}$ to achieve domain adaptation on models’ output between the source domain and the target domain, which is achieved by means of the segmentation discriminator $D_c$.

#### Style Loss. To help the encoder module $E_c$ extract style-independent features, $L_{style}$ also utilizes the original GAN loss \cite{11} to force the style information on the source domain $S_{seg}$ close that on the target domain $S_{te}$.

The loss function during domain adaptation training is summarized as follows

$$L = \lambda_{seg} L_{seg} + \lambda_{adv,seg} L_{adv,seg} + \lambda_{style} L_{style}, \quad (5)$$

where $\lambda$ play a trade-off among these three terms.

During the SSL process, similar with $L_{seg}$, Self-supervised Loss $L_{ssl}$ also utilizes cross entropy function to make the prediction on the target domain $\hat{y}_t \in \mathbb{R}^{H_t \times W_t \times C}$ as close as possible to pseudo labels $y_t \in \mathbb{R}^{H_t \times W_t \times C}$:

$$L_{ssl} = -\frac{1}{H_t \times W_t} \sum_{h=1}^{H_t} \sum_{w=1}^{W_t} \sum_{c=1}^{C} m_t^{hwc} \hat{y}_t^{hwc} \log \hat{y}_t^{hwc}. \quad (6)$$

### 3. EXPERIMENTAL RESULTS

Here we evaluate our model on “GTA5 to Cityscapes” task.

#### 3.1. Datasets

GTA5 \cite{12} includes 24966 synthetic images collected from the game engine. GTA5 have 19-category pixel-accurate annotations compatible with target domain Cityscapes \cite{13}.

Cityscapes \cite{13} is collected from streetscapes in 50 different Germany cities includes training set with 2975 images, validation set with 500 images, testing set with 1525 images. The former two sets contain pixel-wise semantic label maps, while the annotations of testing set are missing. To validate the performance of our model, during testing phase, we use validation set instead of testing set.

#### 3.2. Network Architectures and Implementation Details.

The whole framework of our model is shown in Fig. 2. The encoder module follows DeepLab V2 \cite{14} using ResNet101 \cite{15} as backbone. The parameters are tuned based on weights pretrained on ImageNet \cite{16}. The discriminator $D_f$ for output-based domain adaptation applies PatchGAN \cite{17} to output a 16x downsampled confidence probability map relative to the input semantic segmentation map. The style discriminator $D_j$ also utilizes PatchGAN \cite{17}, but it applies four 1-D convolutional layers with kernel size of 4. All modules are parameter-shared except the style discriminators (i.e., $D_j^1$ and $D_j^2$), and segmentation discriminator $D_c$.

Note that the all game-synthetic source domain images (GTA5 datasets) are firstly translated by CycleGAN \cite{18} module of BDL model \cite{5}. SGD optimizer with momentum $= 0.9$ is used to train encoder $E_c$ and decoder modules, where encoder $E_c$ adopts learning rate $lr = 2.5 \times 10^{-4}$, the decoder adopts $lr = 2.5 \times 10^{-3}$. For style discriminator $D_j$ and segmentation discriminator, Adam optimizer is utilized with $\beta = (0.9, 0.99)$ and $lr = 1 \times 10^{-4}$. In addition, “poly” policy for learning rate update with $maxstep = 250,000$ and $power = 0.9$ is introduced to encoder $E_c$ and decoder. $\lambda_{seg}, \lambda_{adv,seg}, \lambda_{style}$ in Equation (5) are set $1, 1 \times 10^{-3}, 1 \times 10^{-3}$, respectively. Style discriminator $D_j$ and segmentation discriminator $D_c$ utilize exponential decay policy to
update \( lr \), where \( \text{decay\_rate} = 0.1, \text{decay\_steps} = 50000 \).

Two rounds of SSL are applied in our experiments.

### 3.3. Results

**Quantitative results:** The results of different related baselines are shown in Table 1. Consistent with previous work, mIoU metric on 19 specific categories is adopted, where the best result on each category is highlighted in bold. Our model has a gain of 1.7 in overall mIoU rather than the state-of-the-art BDL. In addition, compared to another category-balanced SSL model CBST, our model brings +3.2% mIoU improvement, which demonstrates the superiority of our proposed pseudo labeling method.

![Fig. 3](image)

(a) Image (b) GT (c) BDL(gta2cs) (d) Ours(gta2cs)

**Fig. 3:** Qualitative comparisons. From left to right: (a) Original Cityscape images, (b) Ground truth, (c) BDL on “GTA5 to Cityscapes”, (d) Ours on “GTA5 to Cityscapes”.

### Table 2: Ablation study on SSL and style constraints.

| Method     | model | sidewalk | building | wall | fence | pole | light | 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| 38.6   | 30.7  | 84.8| 34.2  | 43.4| 0.4   | 28.3      | 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| 38.9   | 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 | 25.3  | 28.3 | 35.5  | 24.2 | 83.6       | 27.4    | 74.2| 38.6   | 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  | 35.0 | 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 |

**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)    | 45.3 |

**Qualitative results:** Some segmentation examples are shown in Fig. 3. It can be clearly observed that our method makes less visually obvious prediction errors than BDL.

### 4. CONCLUSION

In this paper, we proposed a style gap bridging mechanism and category-adaptive threshold method for SSL on cross-domain semantic segmentation task. The former utilizes adversarial training to narrow the gaps of style information. The latter makes the use of prior semantic distributions to dynamically choose thresholds for self-supervised training on the target domain images, instead of applying fixed thresholds. A series of experiments have shown the effectiveness and superiority of our proposed model.

### 5. ACKNOWLEDGEMENT

This work is supported in part by the National Key R&D Program of China under Grant 2018YFA0701601, and in part by the fellowship of China National Postdoctoral Program for Innovative Talents (BX20200194).
References

[1] M. D. Zeiler, G. W. Taylor, and R. Fergus, “Adaptive deconvolutional networks for mid and high level feature learning,” in 2011 International Conference on Computer Vision. IEEE, 2011, pp. 2018–2025.

[2] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” arXiv preprint arXiv:1508.06576, 2015.

[3] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 289–305.

[4] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 2517–2526.

[5] M. Li, C. Ye, and W. Li, “High-resolution network for photorealistic style transfer,” arXiv preprint arXiv:2008.07514, 2020.

[6] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, “Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6936–6945.

[7] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” CoRR, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[9] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “Cycada: Cycle-consistent adversarial domain adaptation,” in International conference on machine learning. PMLR, 2018, pp. 1989–1998.

[10] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis, “Dean: Dual channel-wise alignment networks for unsupervised scene adaptation,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 518–534.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.