Cross-Domain Semantic Segmentation of Urban Scenes via Multi-Level Feature Alignment

Bin Zhang, Shengjie Zhao, Rongqing Zhang

Key Laboratory of Embedded System and Service Computing, Ministry of Education

School of Software Engineering,

Tongji University, Shanghai, 200092, China

Outline

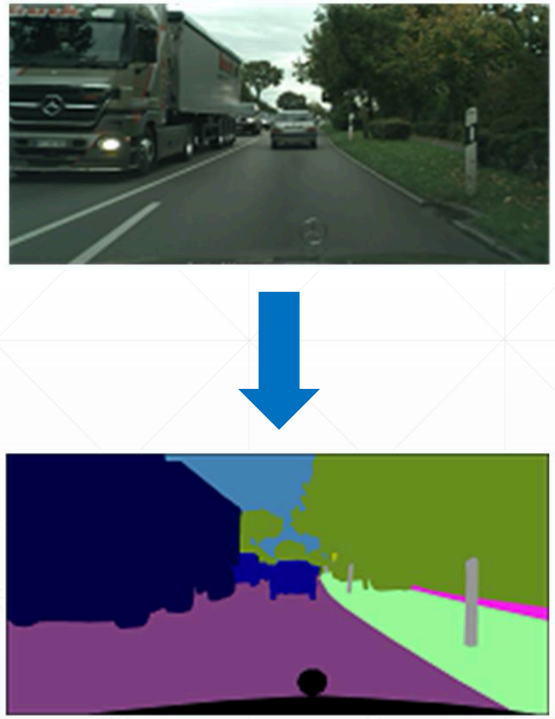
1. Background
2. The Proposed Method
3. Experimental Results
4. Conclusion

Background

In recent years, semantic segmentation has achieved great success due to the development of large-scale dataset

- **Advantage and Challenge of Semantic Segmentation**

Advantage	• This technology has the potential to provide pixel-wise understanding of the scene
Challenge	• The main challenge is the time-consuming process to collect and label the training dataset



Therefore, how to fully utilize the synthetic dataset to improve the model performance in real-world scene is of great importance.

Background

Domain adaptation is a representative method in transfer learning

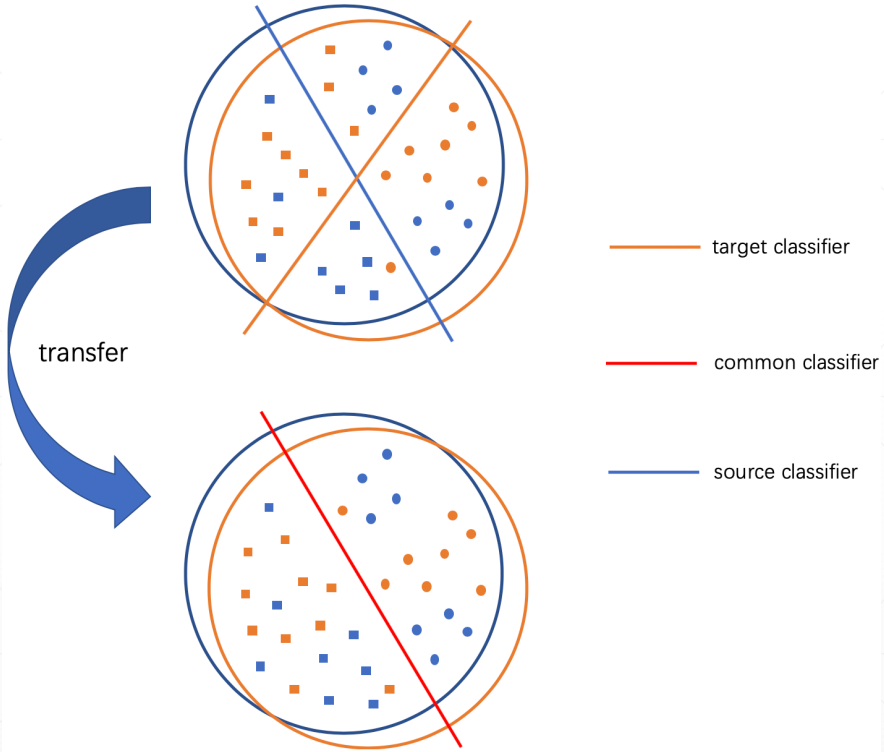
- **Advantage and Main Idea of Domain Adaptation**

Advantage

- Domain adaptation can reduce the retraining process of the model in the target domain

Main Idea

- The main idea is to transfer the model trained on source domain dataset to target domain



Background

The existing domain adaptation techniques can be classified into supervised domain adaptation and unsupervised domain adaptation.

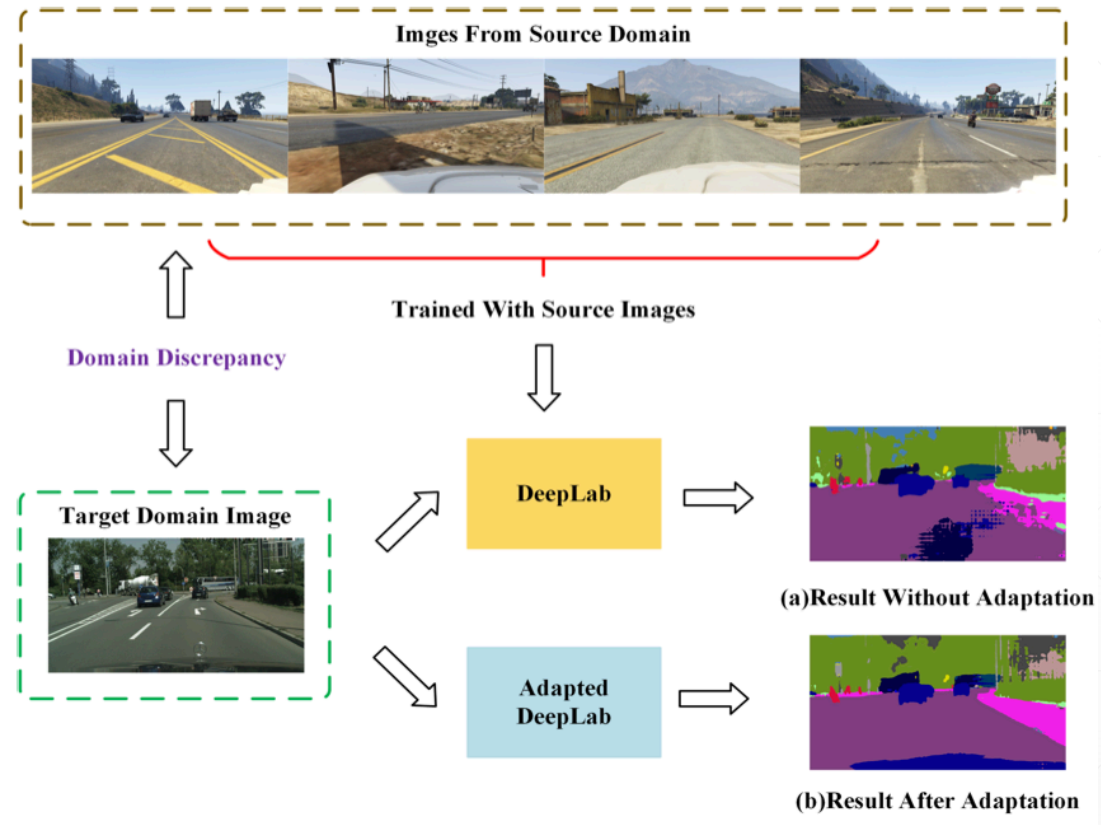
- **Supervised Domain Adaptation**

  Supervised domain adaptation assumes that both the source and target domain samples have the corresponding label

- **Unsupervised Domain Adaptation**

  In the unsupervised domain adaptation setting, the source domain has a large number of annotated training examples while the target domain only have unlabeled data

The Proposed Method

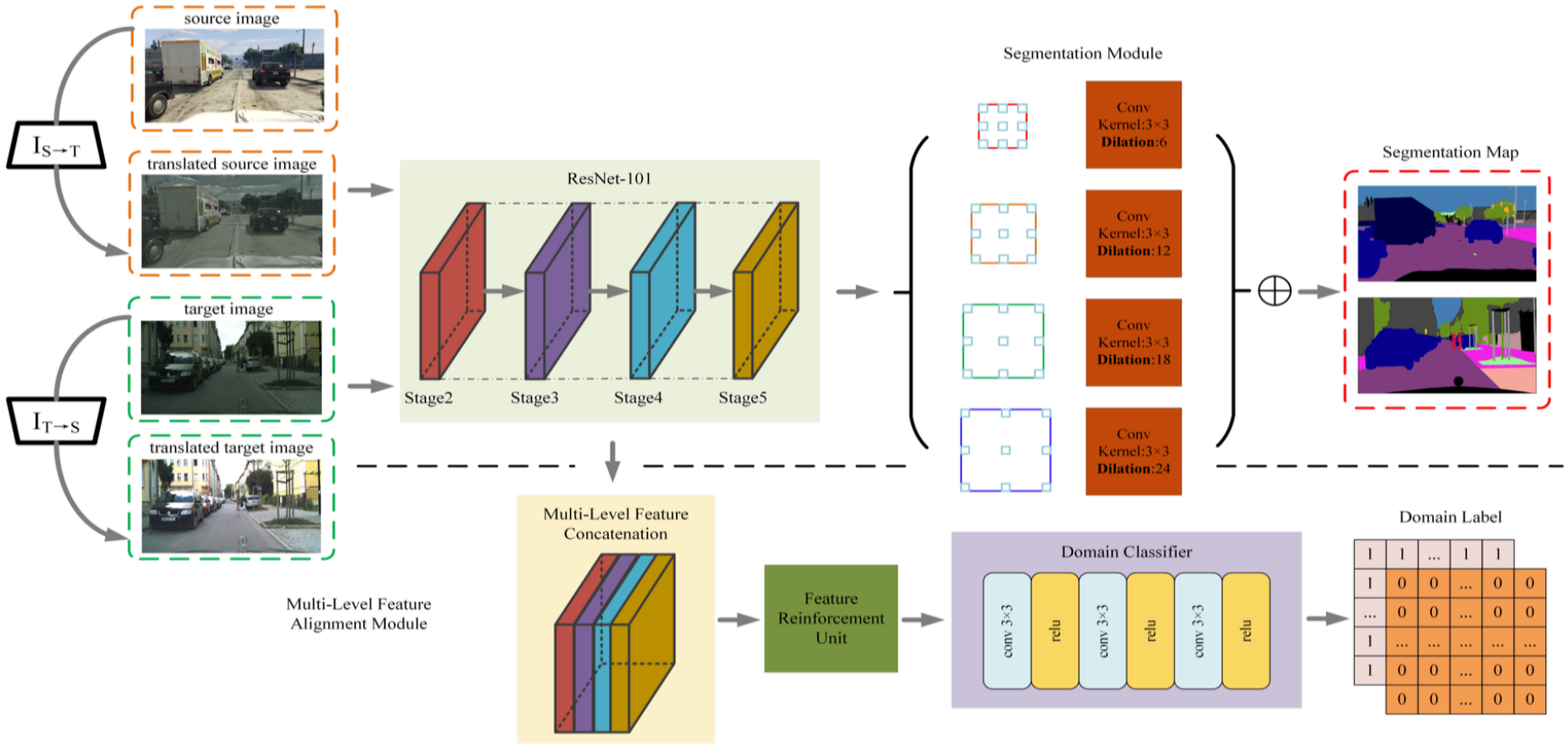


- Comparison between the segmentation results.

(a) Directly applying DeepLab model trained on source images without modification.

(b) Using our multi-level feature alignment method.

The Proposed Method



The Proposed Method

Image-to-Image Translation Network

- By mapping the source image into target domain, we enable our model to learn the segmentation task on labeled source data with target style.

- Two image translators learn to map samples across different domains while two adversarial discriminators try to discriminate them

$$\mathcal{L}_{adv}^{img}\left(I_{S\to T}, D_T\right) = \mathbb{E}_{x\sim\mathcal{X}^t}\left[\log D_T(x)\right] + \mathbb{E}_{x\sim\mathcal{X}^s}\left[\log\left(1 - D_T\left(I_{S\to T}(x)\right)\right)\right]$$

$$\mathcal{L}_{adv}^{img}\left(I_{T\to S}, D_S\right) = \mathbb{E}_{x\sim\mathcal{X}^s}\left[\log D_S(x)\right] + \mathbb{E}_{x\sim\mathcal{X}^t}\left[\log\left(1 - D_S\left(I_{T\to S}(x)\right)\right)\right]$$

$$\mathcal{L}_{cyc}\left(I_{S\to T}, I_{T\to S}\right) = \mathbb{E}_{x\sim\mathcal{X}^s}\left[\left\|I_{T\to S}\left(I_{S\to T}(x)\right) - x\right\|_1\right] + \mathbb{E}_{x\sim\mathcal{X}^t}\left[\left\|I_{S\to T}\left(I_{T\to S}(x)\right) - x\right\|_1\right]$$

The Proposed Method

Semantic Segmentation Network

- We utilize the DeepLab-v2 network with pretrained ResNet-101 backbone as our base model. We discard the last fully connected layer and modify the strides of the last two convolution layers to 1.

$$\mathcal{L}_{seg} = \mathbb{E}_{(x,y)\sim\mathcal{X}^s} - \sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y^{(h,w,c)} \log p^{(h,w,c)}$$

Multi-level Feature Alignment

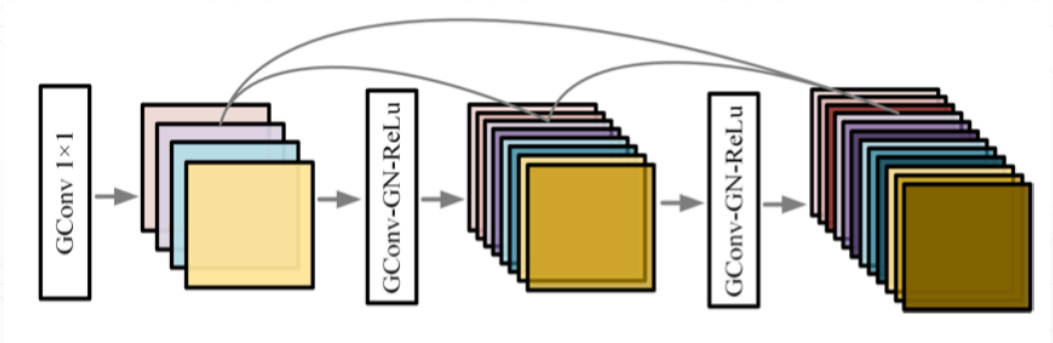
- First, we concatenate the multi-stage feature and obtain the multi-level concatenated feature maps. Then a classifier tries to discriminate the multi-level feature representation

$$\mathcal{L}_d = \mathbb{E}_{x\sim\mathcal{X}^s}\left[\log D\left(F_m^x\right)\right] + \mathbb{E}_{x\sim\mathcal{X}^t}\left[\log\left(1 - D\left(F_m^x\right)\right)\right]$$

$$D\left(F_m^x\right) = \begin{cases} 0, & \text{if } x \sim \mathcal{X}^s \\ 1, & \text{otherwise.} \end{cases}$$

The Proposed Method

- Detailed structure of the feature reinforcement unit. GConv stands for group convolution and GN represents group normalization.



Algorithm 1 Training procedure of our network

Input: source dataset $\mathcal{X}^s$, target dataset $\mathcal{X}^t$, iteration number max_iter, modulating factor $\lambda = 0.001$.

Initialization: Initialize the domain classifier D and the two image translators from scratch. The segmentation network G is pretrained on the ImageNet dataset.

train the image-to-image translation networks $I_{S\to T}$ and $I_{T\to S}$ by optimizing Eqn. (4)
repeat
  $(x^s, y^s) \leftarrow$ sample source image from $\mathcal{X}^s$
  $x^t \leftarrow$ sample target image from $\mathcal{X}^t$
  obtain translated source image $\overline{x^s} = I_{S\to T}(x^s)$
  generate feature maps for both $\overline{x^s}$ and x^t
  $\mathcal{L}_{seg} \leftarrow$ compute the segmentation loss for source image by Eqn. (5)
  $G \leftarrow \min \mathcal{L}_{seg} + \lambda \mathcal{L}_{adv}$
  $D \leftarrow \min \mathcal{L}_d$
until max_iter

Experimental Results

- **Datasets and Evaluation Metric**

  - We verify the performance of our proposed approach on the GTA5 → Cityscapes and SYNTHIA → Cityscapes domain adaptation tasks. Cityscapes is a large-scale dataset to evaluate the accuracy of semantic segmentation models, which covers the urban scenes of several European countries. It is split into a training set with 2,975 samples, a testing set with 1,525 samples, and a validation set with 500 samples. GTA5 dataset contains 24,966 high-definition images collected from a contemporary computer game called Grand Theft Auto V. The dataset is automatically annotated into 19 categories, which are consistent with the Cityscapes dataset. As for the evaluation metric, we choose the commonly adopted Intersection over Union (IoU) for fair comparison:

$$IoU = \frac{TP}{TP + FP + FN}$$

Experimental Results

TABLE I
THE COMPARISON RESULTS BETWEEN BASELINE APPROACHES AND OURS FROM GTA5 TO CITYSCAPES.

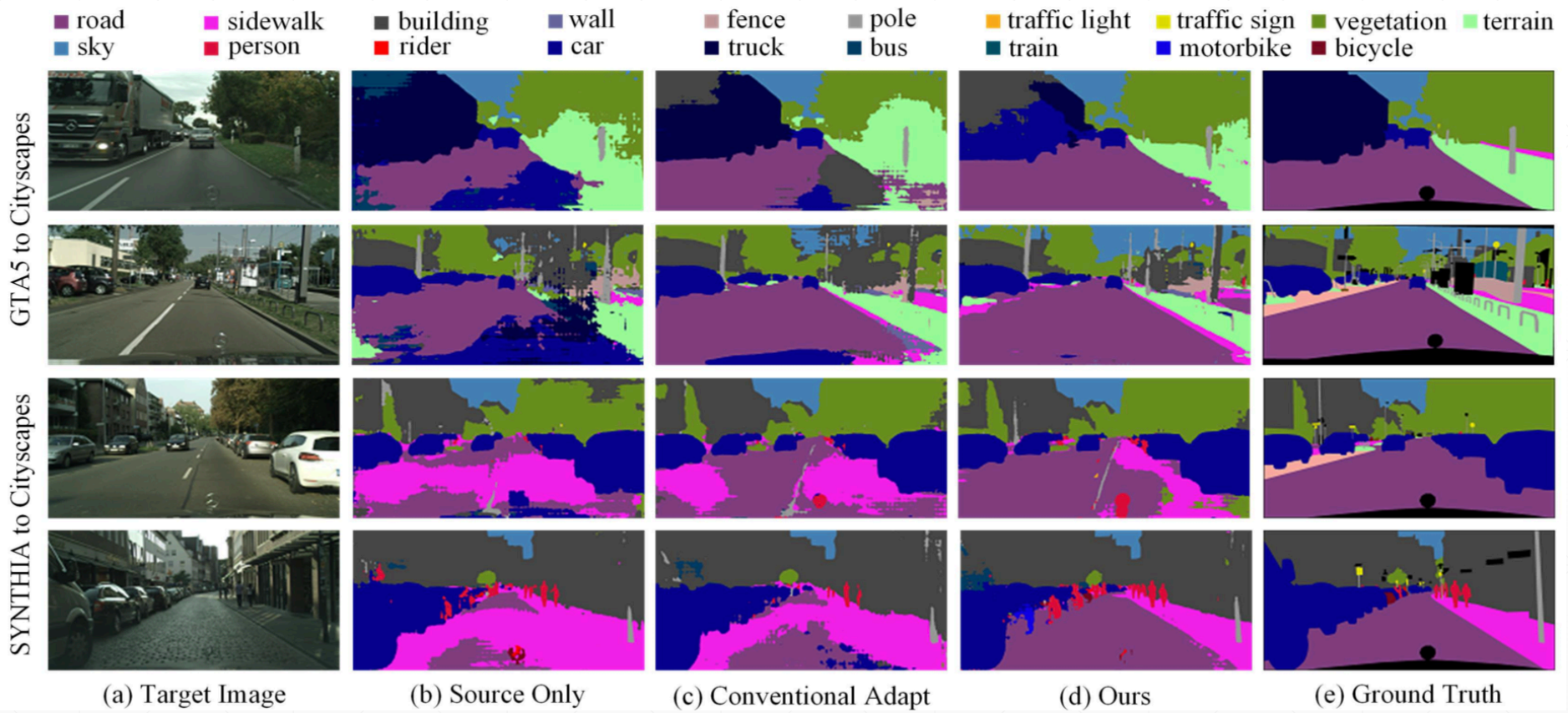
Methods	Backbone	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	motorbike	bike	mIoU
FCN WId [13]	VGG-16	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
TGCF-DA [20]	VGG-16	90.2	**51.5**	81.1	15.0	10.7	**37.5**	35.2	28.9	**84.1**	32.7	75.9	**62.7**	19.9	82.6	22.9	28.3	0.0	19.3	12.0	42.5
ROAD [21]	VGG-16	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	**9.4**	8.3	0.0	35.9
DLOW [22]	ResNet-101	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	0.0	22.9	**34.5**	42.3
CLAN [23]	ResNet-101	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
AdvEnt [24]	ResNet-101	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
Ours	ResNet-101	**91.8**	47.1	**85.6**	**29.5**	**29.3**	35.4	**36.8**	**33.2**	81.5	35.1	**82.1**	62.1	**30.6**	79.0	22.6	33.4	4.3	**33.9**	20.1	**45.9**

TABLE II
THE COMPARISON RESULTS BETWEEN BASELINE APPROACHES AND OURS FROM SYNTHIA TO CITYSCAPES.

Methods	Backbone	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	sky	person	rider	car	bus	motorbike	bike	mIoU
FCN WId [13]	VGG-16	11.5	19.6	30.8	4.4	0.0	20.3	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	20.2
TGCF-DA [20]	VGG-16	**90.1**	**48.6**	80.7	2.2	0.2	27.2	3.2	14.3	82.1	78.4	**54.4**	16.4	**82.5**	12.3	1.7	21.8	38.5
ROAD [21]	VGG-16	77.7	30.0	77.5	**9.6**	0.3	25.8	10.3	**15.6**	77.6	79.8	44.5	16.6	67.8	14.5	7.0	23.8	36.2
CLAN [23]	ResNet-101	81.3	37.0	80.1	–	–	–	**16.1**	13.7	78.2	81.5	53.4	**21.2**	73.0	**32.9**	**22.6**	30.7	–
AdaptSegNet [14]	ResNet-101	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**	–
Ours	ResNet-101	86.6	43.3	**81.5**	4.3	**1.1**	**30.2**	10.2	8.7	**85.3**	**82.2**	48.8	16.9	73.7	30.4	11.3	29.5	**40.2**

- Our method outperforms other methods by a large margin in segmenting "road", "building", "wall", "fence", "sky", "rider", and "motorbike" categories.

Experimental Results



Conclusions

- **Contributions**

1. we propose a novel multi-level feature alignment method for cross-domain semantic segmentation.

2. Our proposed MLFA provides a novel perspective of insight by incorporating the content and style alignment module.

3. The experimental results demonstrate that MLFA outperforms most current state-of-the-art unsupervised domain adaptation methods.

References

THANK YOU!

Bin Zhang, Shengjie Zhao, Rongqing Zhang

Key Laboratory of Embedded System and Service Computing, Ministry of Education

School of Software Engineering,

Tongji University, Shanghai, 200092, China