NEURAL INFORMATION PROCESSING SYSTEMS

When Visual Prompt Tuning Meets Source-Free Domain Adaptive Semantic Segmentation

Xinhong Ma, Yiming Wang, Hao Liu, Tianyu Guo, Yunhe Wang

Huawei Noah's Ark Lab

Mindspore: https://gitee.com/mindspore/models/tree/master/research/cv/uni-uvpt
Pytorch: https://github.com/huawei-noah/noah-research/tree/master/uni-uvpt

Source-Free Domain Adaptive Semantic Segmentation

- **Definition:**

  Adapting a pretrained source model to the unlabeled target domain without accessing the private source data.

- **Limitations:**

  Previous methods usually finetune the entire network, which suffers from expensive parameter tuning.

How to achieve parameter-efficient adaption?

Prompt tuning may make a difference!

Prompt Tuning

- **Definition:**

  Designing a trainable lightweight block as a supplementary input (prompt) for a frozen model, which guides or directs the generalization of representations to achieve desirable performances.

- **Limitations of existing visual prompt tuning methods:**

  a) *The learned visual prompts are unreasonable.*

  b) *Lacking methods addressing downstream tasks without sufficient labeled data, i.e., unsupervised visual prompt tuning.*

We propose a Universal Unsupervised Visual Prompt Tuning (Uni-UVPT) framework for source-free domain adaptive semantic segmentation.

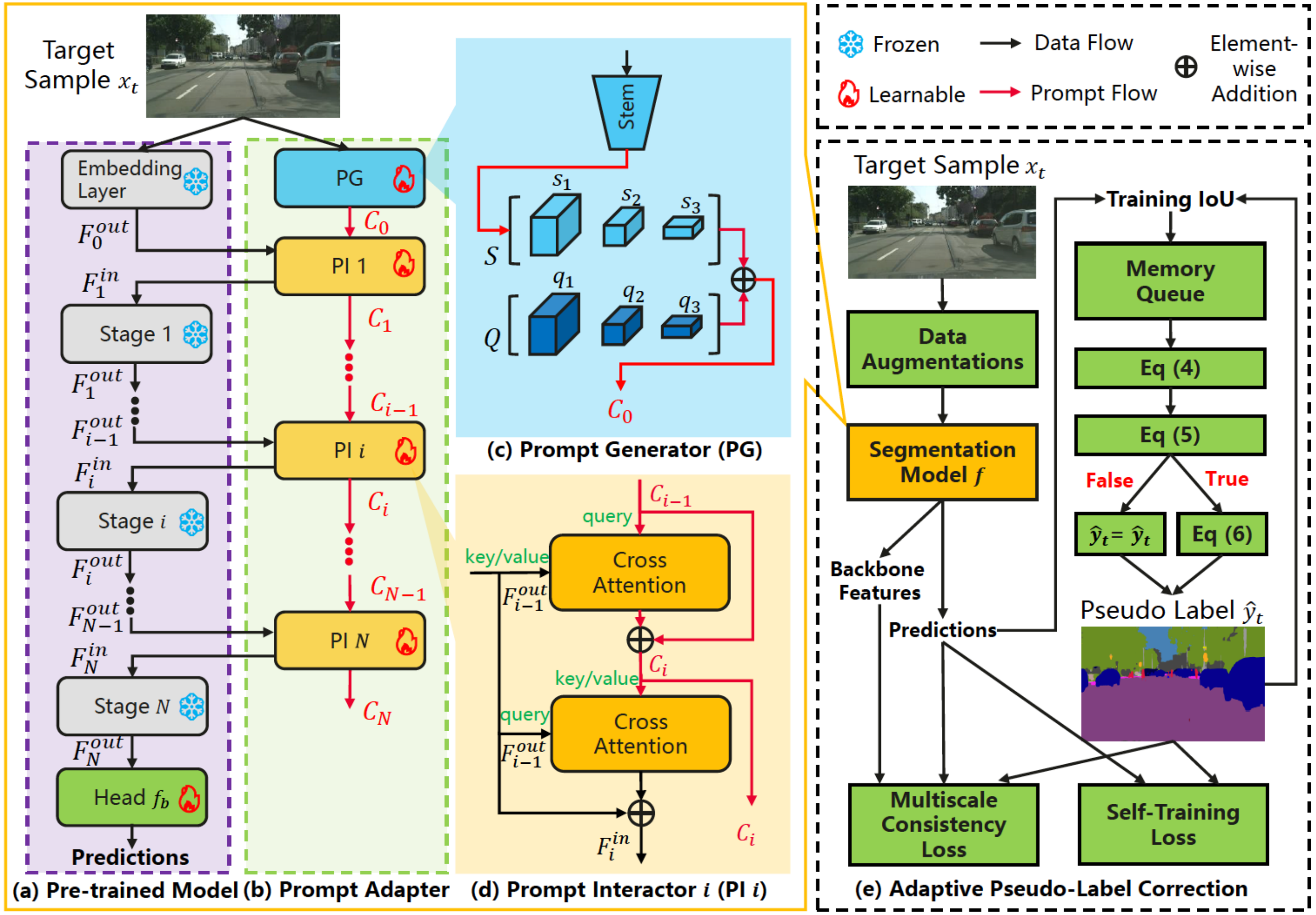
Universal Unsupervised Visual Prompt Tuning

➢ Prompt Adapter:

- Generating informative visual prompts.
- Improving the generalization of target features.

➢ Adaptive Pseudo-Label Correction

- Learning visual prompts with massive unlabeled target data.
- Enhancing visual prompt's capacity for spatial perturbations.



Prompt Adapter

➢ Prompt Generator:

- Stem (S): a convolutional network for capturing multiscale spatial information.
- Level Embedding (Q): trainable vectors for learning task-shared knowledge.

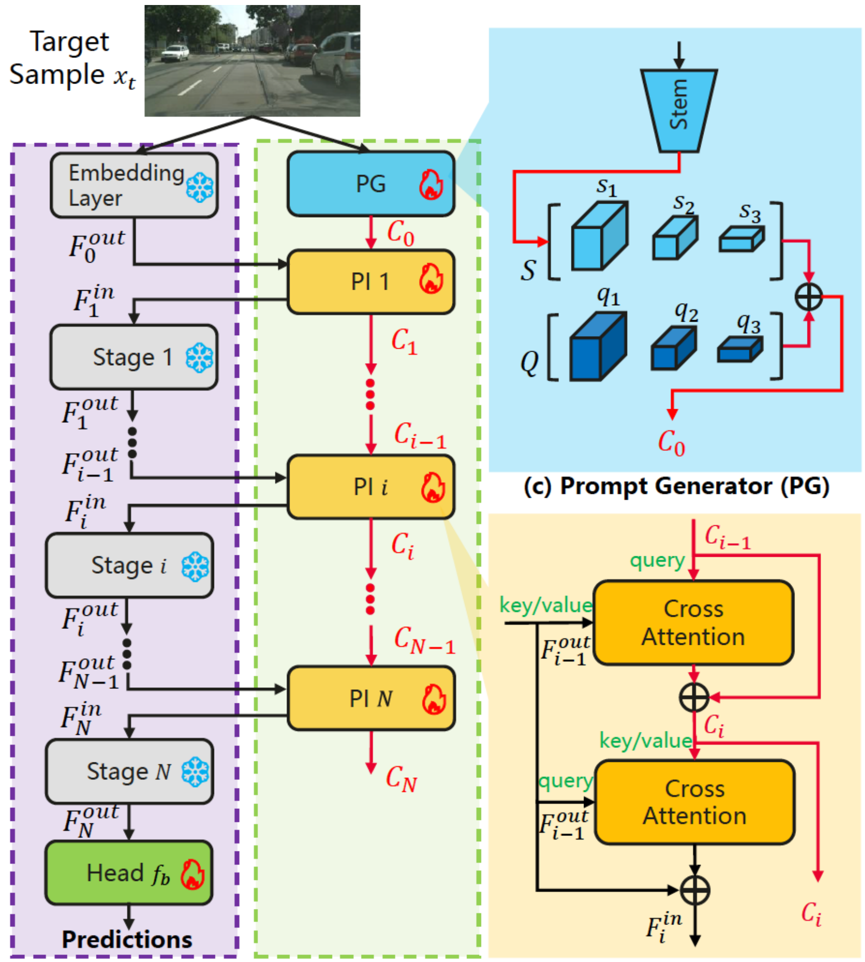
➢ Prompt Interactor

- Injecting pretrained knowledge into prompts:

$$C_i = C_{i-1} + \mathrm{Attention}(\mathrm{norm}(C_{i-1}), \mathrm{norm}(F_{i-1}^{out}))$$

- Generating adapted features with refined prompts:

$$F_i^{in} = F_{i-1}^{out} + \gamma_i \cdot \mathrm{Attention}(\mathrm{norm}(F_{i-1}^{out}), \mathrm{norm}(C_i))$$

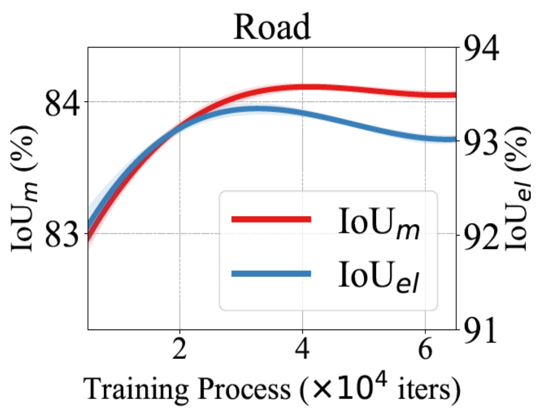


a) Pre-trained Model (b) Prompt Adapter (c) Prompt Generator (PG) (d) Prompt Interactor i (PI i)

Adaptive Pseudo-Label Correction

Early-learning phenomenon

- Deep models tend to first fit data with correct pseudo labels during early-learning phase, before eventually memorizing instances with incorrect/noisy pseudo labels.
- The performance deceleration of IoU_m indicates whether overfitting noisy pseudo labels.



Correcting pseudo-labels at suitable moments

- Fitting the training IoU using the least squares:

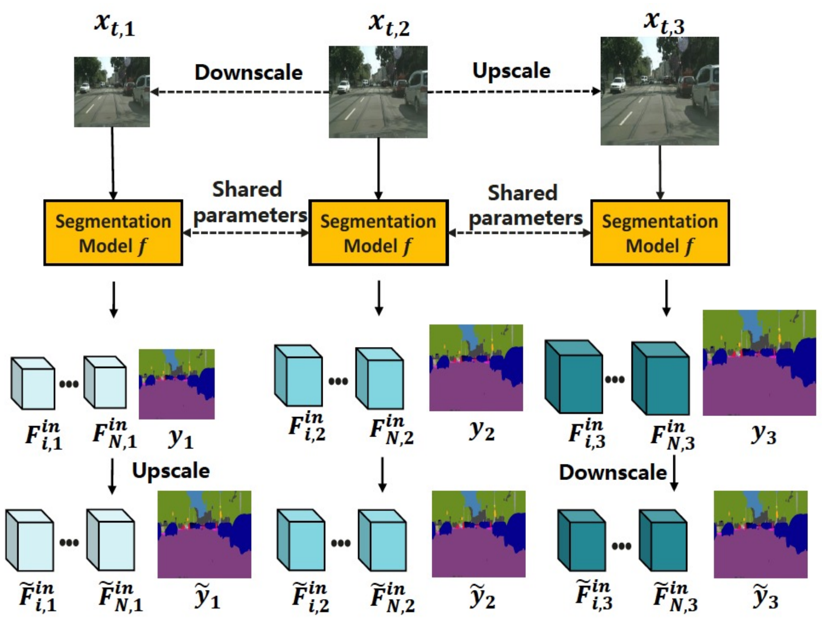
$$g(t) = at^3 + bt^2 + ct + d$$

- The correction for each category are performed when the condition is satisfied:

$$\frac{|g'(t_0) - g'(t)|}{|g'(t_0)|} > \tau$$

- The correct pseudo label can be obtained by averaging predictions of multiple rescaled input samples:

$$\hat{y}_t = \frac{1}{m}\sum_{k=1}^{m} \tilde{y}_k$$



Multiscale consistency loss

$$\mathcal{L}_{mc} = \underbrace{\alpha\,\mathbb{E}_{x_t\sim\mathcal{D}_t}\left[\sum_{i=1}^{N}\frac{1}{m}\sum_{k=1}^{m}\|\tilde{F}^{in}_{i,k} - \hat{F}^{in}_{i}\|_2^2\right]}_{\text{feature consistency } \mathcal{L}_{fc}} \underbrace{+\beta\,\mathbb{E}_{x_t\sim\mathcal{D}_t}\left[-\frac{1}{m}\sum_{k=1}^{m}\mathrm{KL}\left(\tilde{y}_k \,\|\, \hat{y}_t\right)\right]}_{\text{prediction consistency } \mathcal{L}_{pc}}$$

Experiments: Comparative Results

Table 1: Quantitative evaluations on GTA5 → Cityscapes and SYNTHIA → Cityscapes tasks. Different segmentation architectures: F (FCN8s VGG-16), D (DeepLabv2 ResNet-101), S (Swin-B), M (MiT-B5). FB: whether the backbone is frozen. Params (M): number of trainable parameters. **Bold**: the best results based on different source pre-trained models. (+x.x): mIoU gains over the corresponding source pre-trained models where the best are in red. Underline: the state-of-the-art results. The full table with per-class IoUs is available in the appendices.

Methods	Arch	FB	Params (M)	GTA5 → Cityscapes	SYNTHIA → Cityscapes	
				$mIoU_{19}(\%)$	$mIoU_{16}(\%)$	$mIoU_{13}(\%)$
SFDA [30]	F	✗	-	35.8	-	-
GtA [19]	F	✗	134.5	45.9	41.3	48.9
URMA [14]	D	✗	47.4	45.1	39.6	45.0
SRDA [5]	D	✗	-	45.8	-	-
SFUDA [46]	D	✗	-	49.4	-	51.9
BDT [20]	D	✗	43.8	52.6	-	56.7
GtA [19]	D	✗	43.8	53.4	52.0	60.1
Standard Single Source	S	✗	90.7	50.5	44.6	49.8
CPSL [24]	S	✓	3.9	51.1 (+0.6)	46.4 (+1.8)	52.3 (+2.5)
VPT [18] + ELR [47]	S	✓	7.0	53.5 (+2.0)	47.7 (+3.1)	53.2 (+3.4)
Ours	S	✓	28.6	**56.2** (+5.7)	**52.6** (+8.0)	**59.4** (+9.6)
Standard Single Source	M	✗	85.2	52.5	48.6	55.0
CPSL [24]	M	✓	3.7	52.5 (+0.0)	50.5 (+1.9)	57.2 (+2.1)
VPT [18] + ELR [47]	M	✓	7.6	54.1 (+1.6)	51.6 (+3.0)	58.0 (+3.0)
Ours	M	✓	12.3	**54.2** (+1.7)	**52.6** (+4.0)	**59.3** (+4.3)
Source-GtA [19]	S	✗	110.4	52.8	48.8	55.0
CPSL [24]	S	✓	3.9	53.5 (+0.7)	49.6 (+0.8)	56.2 (+1.2)
VPT [18] + ELR[47]	S	✓	7.0	55.1 (+2.3)	51.6 (+2.8)	58.2 (+3.2)
GtA [19]	S	✓	23.6	56.1 (+3.3)	52.5 (+3.7)	58.7 (+3.7)
Ours	S	✓	28.6	**56.9** (+4.1)	**53.8** (+5.0)	**60.4** (+5.4)
Source-GtA [19]	M	✗	103.7	53.0	50.0	56.2
CPSL [24]	M	✓	3.7	53.2 (+0.2)	52.2 (+2.2)	58.7 (+2.5)
VPT [18] + ELR [47]	M	✓	7.6	54.4 (+1.4)	53.0 (+3.0)	59.5 (+3.3)
GtA [19]	M	✓	22.3	55.2 (+2.2)	53.6 (+3.6)	59.7 (+3.5)
Ours	M	✓	12.3	**56.1** (+3.1)	**53.8** (+3.8)	**60.1** (+3.9)

Experiments: Ablation Study

Table 2: Ablation study on the prompt adapter. PG, PI and LE respectively denote prompt generator, prompt interactor and level embedding. The performance drop is over our complete approach.

PG		PI	mIoU (%)
Stem	LE		
Multiscale	✓	✓	56.24 (Ours)
Multiscale	✗	✓	55.58 $\downarrow$0.66
Singlescale	✓	✓	55.52 $\downarrow$0.72
✗	✓	✓	55.50 $\downarrow$0.74
Multiscale	✓	PI 1	55.34 $\downarrow$0.90
✗	✗	✗	55.07 $\downarrow$1.17

Table 3: Analysis on pseudo-label strategies.

Methods	mIoU (%)
Ours	56.24
ELR [47]	55.60
Ours + offline	55.47

Table 4: Analysis on consistency loss.

Feature	Prediction	mIoU (%)
✓	✓	56.24 (Ours)
✗	✓	54.26
✓	✗	56.01
✗	✗	53.81

Table 5: Comparative results of different augmentations on GTA5 $\rightarrow$ Cityscapes and SYNTHIA $\rightarrow$ Cityscapes tasks. Different segmentation architectures: S (Swin-B), M (MiT-B5). FB: whether the backbone is frozen. Params (M): number of trainable parameters. (+x.x): mIoU gains over the corresponding source pre-trained models.

Methods	Arch	FB	Params (M)	GTA5 $\rightarrow$ Cityscapes	SYNTHIA $\rightarrow$ Cityscapes
				mIoU_{19}(%)	mIoU_{16}(%)
Ours	S	✓	28.6	56.2 (+5.7)	52.6 (+8.0)
Ours-weather	S	✓	28.6	54.7 (+4.2)	52.9 (+8.3)
Ours	M	✓	12.3	54.2 (+1.7)	52.6 (+4.0)
Ours-weather	M	✓	12.3	54.1 (+1.6)	53.0 (+4.4)
Ours	S	✓	28.6	56.9 (+4.1)	53.8 (+5.0)
Ours-weather	S	✓	28.6	54.1 (+1.6)	53.0 (+4.4)
Ours	M	✓	12.3	56.1 (+3.1)	53.8 (+3.8)
Ours-weather	M	✓	12.3	55.2 (+2.2)	54.5 (+4.5)

Contribution and Conclusion

(1) We first highlight the low-efficiency problem of fine-tuning large-scale backbones in source-free domain adaptive semantic segmentation, and propose a universal unsupervised visual prompt tuning framework for parameter-efficient model adaptation.

(2) A lightweight prompt adapter is introduced to learn reasonable visual prompts and enhance feature generalization in a progressive manner. Cooperatively, a novel adaptive pseudo-label correction strategy is proposed to rectify target pseudo labels at suitable moments and improve the learning capacity of visual prompts.

Thank you!

Any problem, please contact the primary authors:

Xinhong Ma, Yiming Wang, Hao Liu, Tianyu Guo, Yunhe Wang

stefanxinhong@gmail.com