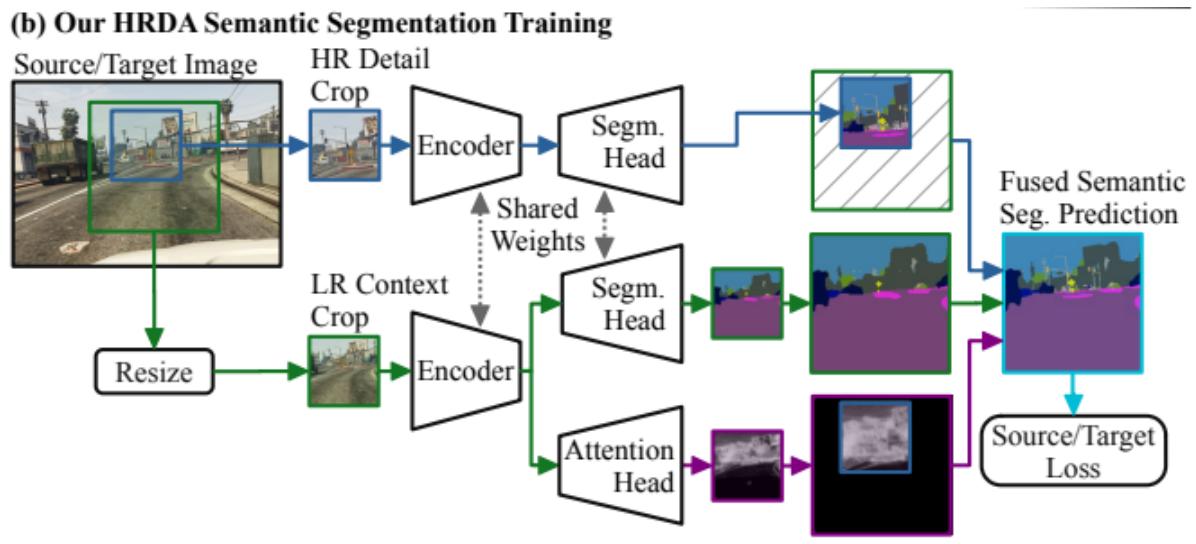
HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation

arxiv link: https://arxiv.org/pdf/2204.13132.pdf

github link: https://github.com/lhoyer/HRDA

presentation link: https://www.youtube.com/watch?v=z9OJdaJ0i24



Overall methodology

Introduction

> 💡 **Unsupervised Domain Adaption (UDA)**
>
> aims to adapt a model trained on the source domain (e.g. synthetic data) to the target domain (e.g. real-world data) without requiring further annotations on the target domain

- **Challenges with UDA**

UDA methods are usually more GPU memory intensive than regular supervised learning as UDA training often requires images from multiple domains, additional networks (e.g. teacher model or domain discriminator), and additional losses, which consume significant additional GPU memory.

→ most UDA semantic segmentation methods so far follow the convention of downscaling images due to GPU memory constraints

▼ Example

Taking Cityscapes as an example, current UDA methods use half the full resolution (i.e. 1024×512 pixels), while most supervised methods use the full resolution (i.e. 2048×1024 pixels). This is one of the key differences in the training setting of UDA and supervised semantic segmentation, possibly contributing to the gap between the state-of-the-art performance of UDA and supervised learning.

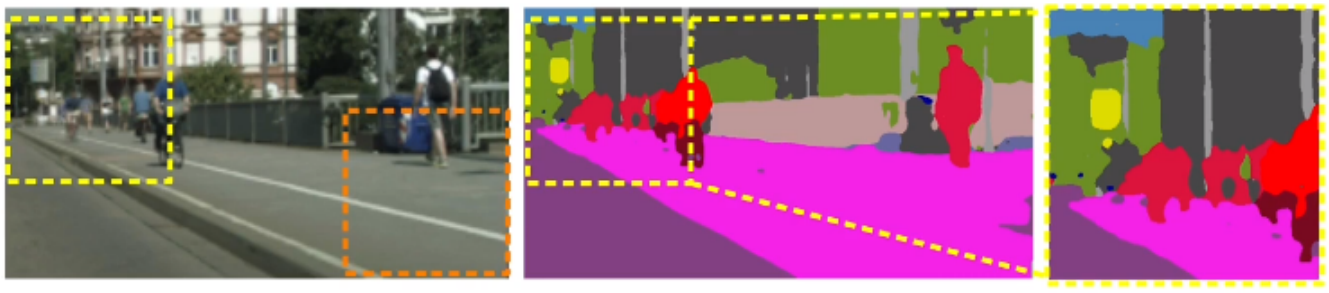
Contributions

1. systematically studying the influence of resolution and crop size
2. exploiting HR inputs for adapting small objects and fine segmentation details

3. applying multi-resolution fusion with a learned scale attention for object-scale-dependent adaptation
4. fusing a nested large LR crop to capture long-range context and small HR crop
   to capture details for memory-efficient UDA training.

Related Works/Motivation

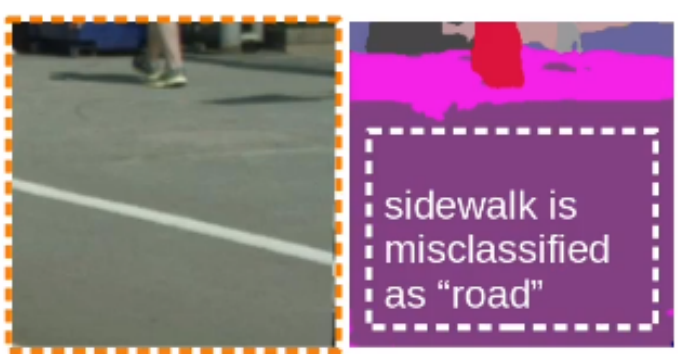
Prev approach #1: Train with down-scaled images



→ Fails to preserve fine details

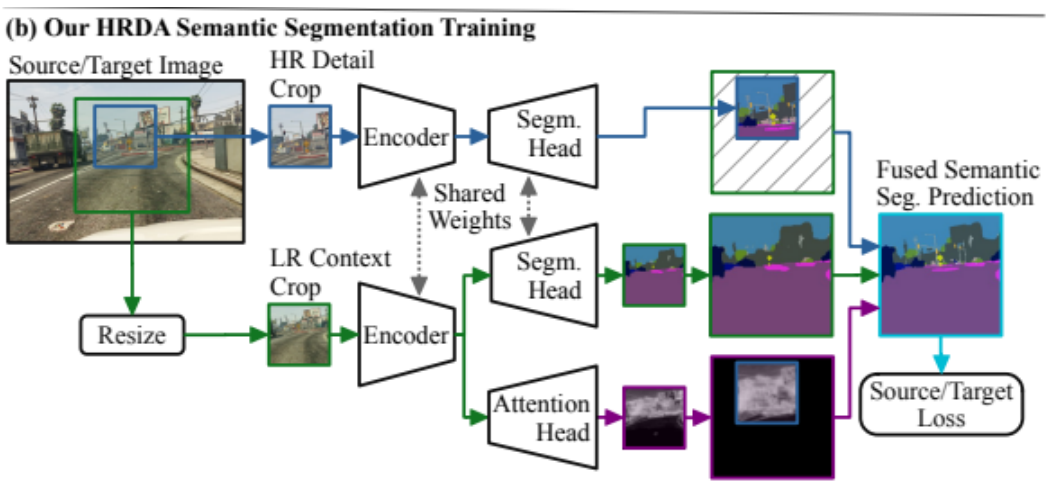
Prev approach #2: Train with high-resolution random crops

→ fine details preserved!

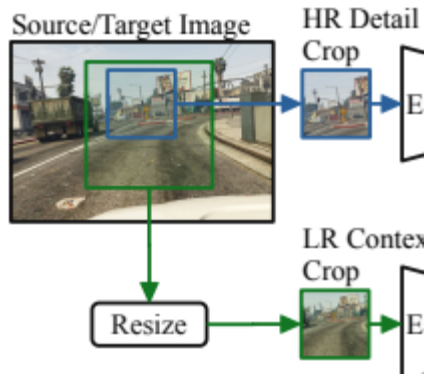


→ fails to capture context!

Method 1: Training



Context and Detail Crop



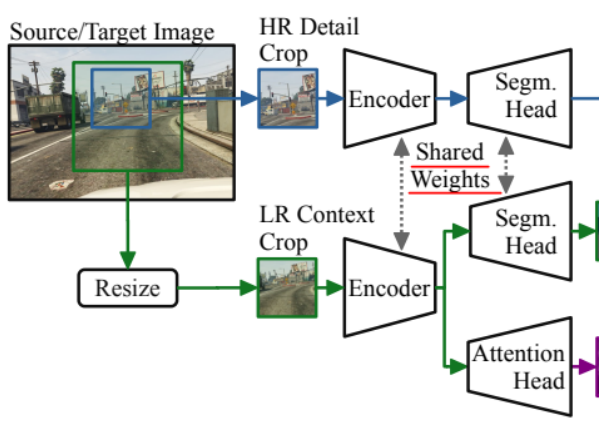
Two types of resolution!

Use both types of resolution!

: small objects and segmentation details are easier to adapt with high-resolution (HR) inputs (contributes to produce fine segmentation borders), while large stuff regions are easier to adapt with low-resolution (LR) inputs.

(most previous works only use LR inputs)

The context crop (LR) covers 4 times more content at half the resolution compared to the detail crop (HR).

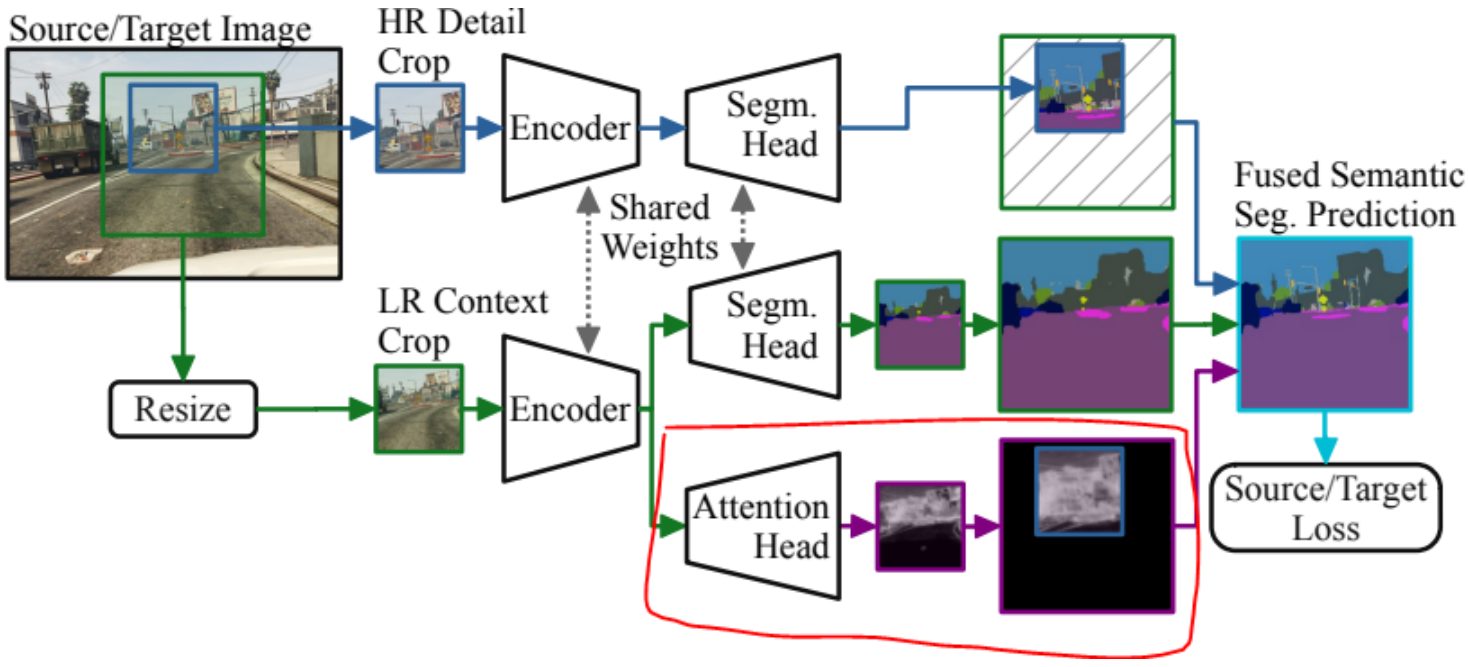


Share weights!

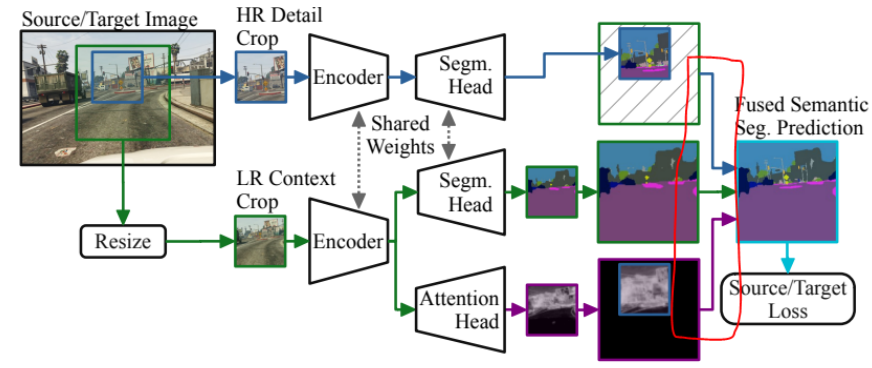
The networks f^E and f^S are shared for both HR and LR inputs. This not only saves memory usage but also increases the robustness of the network against different resolutions.

Multi-Resolution Fusion

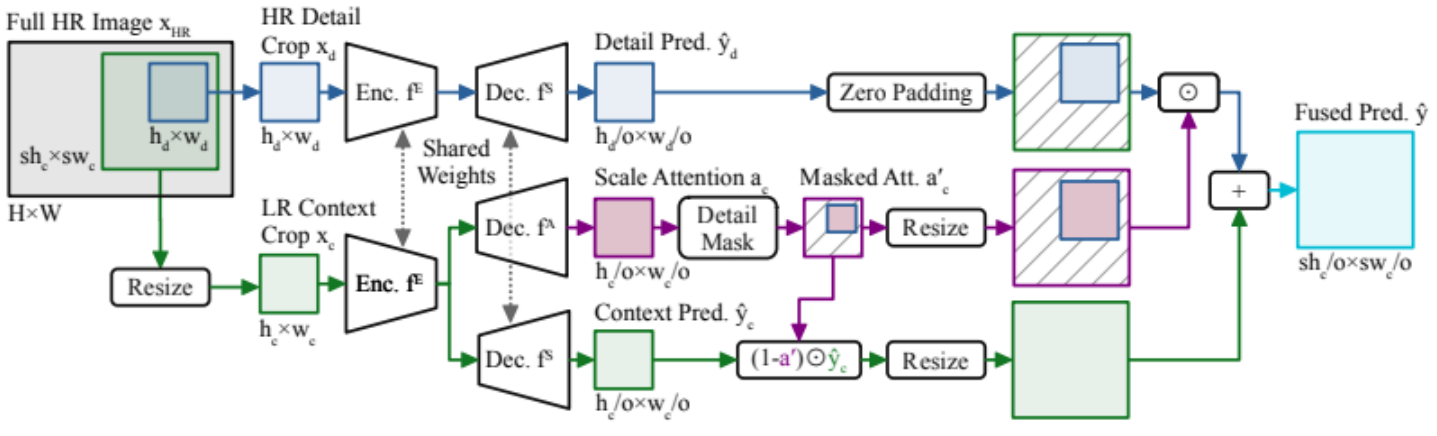
Therefore, we fuse the predictions from both crops using a learned scale attention to predict in which image regions to trust predictions from context and detail crop.



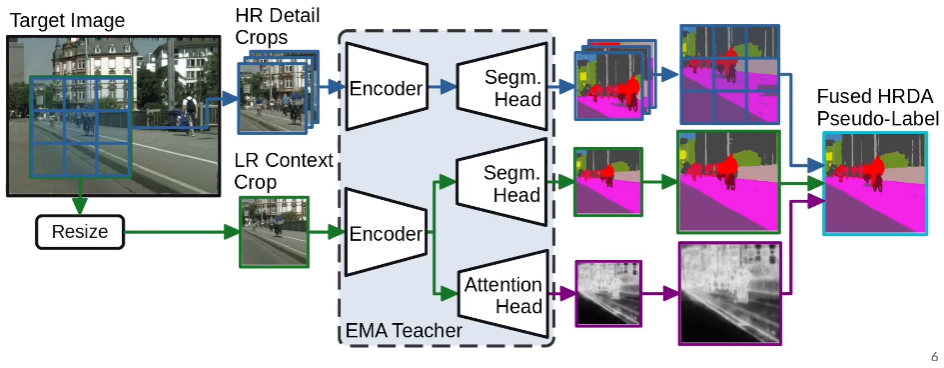
The scale attention decoder f^A learns to predict the scale attention $a_c = \sigma(f^A(f^E(x_c)))$ to weigh the trustworthiness of LR context and HR detail predictions. The sigmoid function σ ensures a weight in [0, 1], where 1 means a focus on the HR detail crop. (The detail crop is aligned with the (upsampled) context crop by padding it with zeros)



The predictions from multiple scales are fused using the attention-weighted sum.



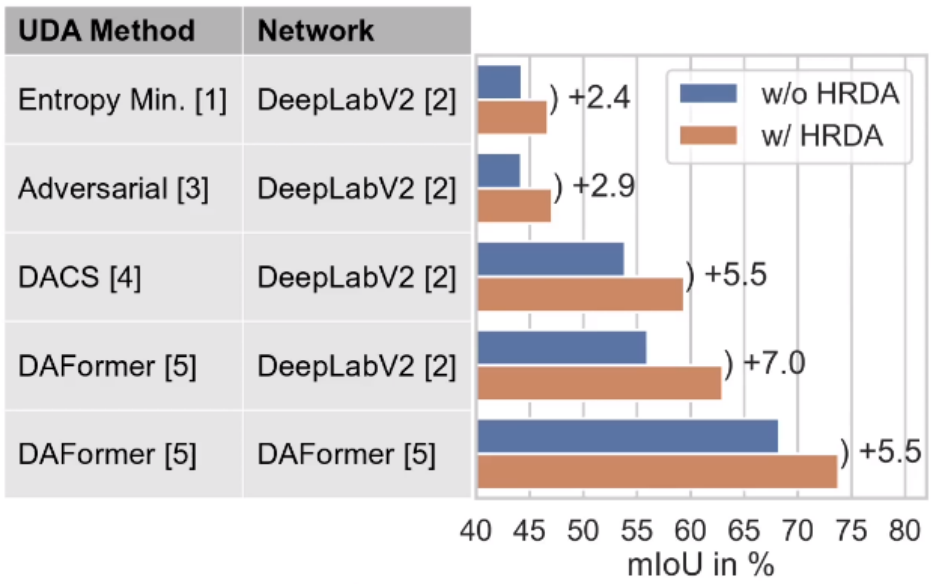
Method 2: (Inference) Pseudo-Label Generation



Use of overlapping sliding window (sliding window를 활용하여 사용할 영역을 지정할 때 이전 영역과 이후 영역에 조금씩 겹치게 끔 함) → multiple views with different contexts are considered for robustness

(나머지는 method 1과 동일)

Experiments



HRDA improves performances across different UDA methods and network architecture!

(= 다른 UDA 방법과 함께 적용하면 성능 향상 & 다른 모델에 적용하기에 용이하고 효과적!)

Comparison with State-of-the-Art UDA Methods

[Quantitative Comparisons]

Table 1. Comparison with previous UDA methods. The results of HRDA are averaged over 3 random seeds. Further methods are shown in the supplement.

	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
GTA5 → Cityscapes																				
CBST [99]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
DACS [63]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
CorDA [71]	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
BAPA [41]	94.4	61.0	88.0	26.8	39.9	38.3	46.1	55.3	87.8	46.1	89.4	68.8	40.0	90.2	60.4	59.0	0.0	45.1	54.2	57.4
ProDA [87]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
DAFormer [29]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
HRDA (Ours)	**96.4**	**74.4**	**91.0**	**61.6**	**51.5**	**57.1**	**63.9**	**69.3**	**91.3**	**48.4**	**94.2**	**79.0**	**52.9**	**93.9**	**84.1**	**85.7**	**75.9**	**63.9**	**67.5**	**73.8**
Synthia → Cityscapes																				
CBST [99]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	–	78.3	60.6	28.3	81.6	–	23.5	–	18.8	39.8	42.6
DACS [63]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
BAPA [41]	91.7	53.8	83.9	22.4	0.8	34.9	30.5	42.8	86.6	–	88.2	66.0	34.1	86.6	–	51.3	–	29.4	50.5	53.3
CorDA [71]	**93.3**	**61.6**	85.3	19.6	5.1	37.8	36.6	42.8	84.9	–	90.4	69.7	41.8	85.6	–	38.4	–	32.6	53.9	55.0
ProDA [87]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	**88.1**	–	84.4	74.2	24.3	88.2	–	51.1	–	40.5	45.6	55.5
DAFormer [29]	84.5	40.7	88.4	41.5	**6.5**	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	87.2	–	53.2	–	53.9	61.7	60.9
HRDA (Ours)	85.2	47.7	**88.8**	**49.5**	4.8	**57.2**	**65.7**	**60.9**	85.3	–	**92.9**	**79.4**	**52.8**	**89.0**	–	**64.7**	–	**63.9**	**64.9**	**65.8**

HRDA improves the IoU of almost all classes across both datasets.

[Qualitative Comparisons]

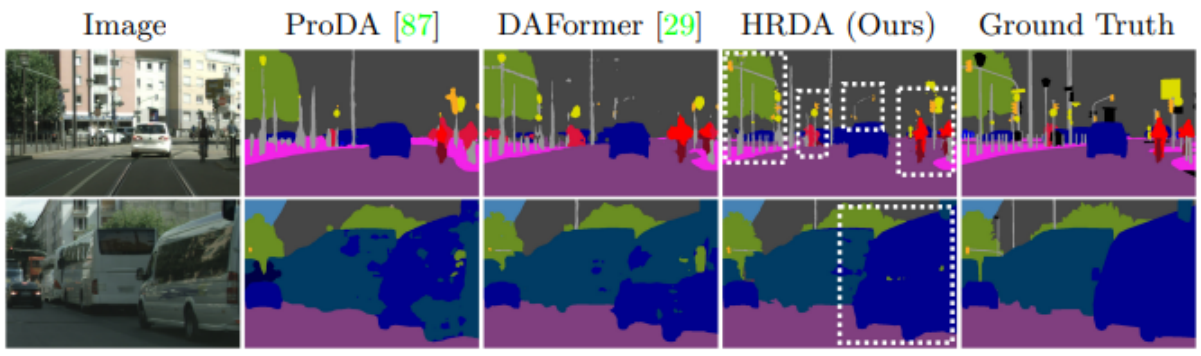
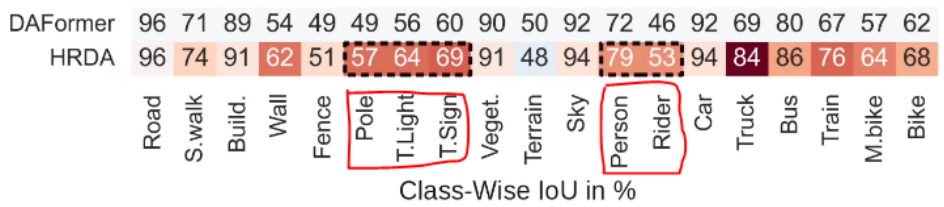
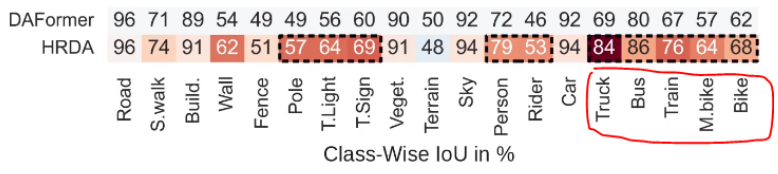


Fig. 3. Qualitative comparison of HRDA with previous methods on GTA→Cityscapes. HRDA improves the segmentation of small classes such as pole, traffic sign, traffic light, and rider as well as large and difficult classes such as bus.

[Performance Analysis]

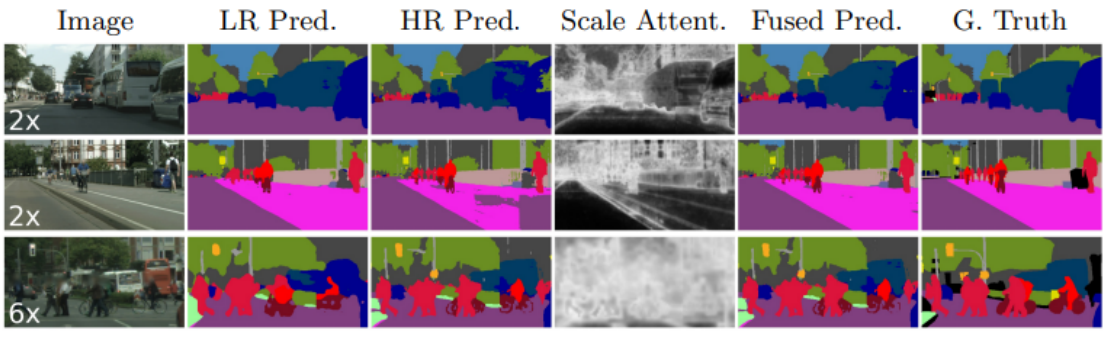


Very effective for small objects (ex. Person, Rider)



Also effective for semantically similar objects (ex. Truck, bus)

HRDA Component Ablations (Qualitative Analysis)



Visual examples of the different predictions and the scale attention of HRDA