Instance-Warp: Saliency Guided Image Warping for Unsupervised Domain Adaptation
Supplementary Material

Shen Zheng* Anurag Ghosh* Srinivasa G. Narasimhan
Carnegie Mellon University
{shenzhen, anuraggh, srinivas}@cs.cmu.edu

A. Synthetic to Real Domain Adaptation

The *Synthetic-to-Real (Sim2Real) Domain Gap is not the focus of this work, as synthetic scene datasets may have suboptimal simulations and inconsistent viewpoints, resulting in unrealistic object-background distributions with non-dominant backgrounds, where our strategy of oversampling objects is less effective.*

However, we do explore how our method impacts performance on the Sim2Real domain gap by considering the GTA [10] and Synthia [11] synthetic datasets. We expect our method to perform better with realistic datasets and less effectively with unrealistic datasets.

As shown in Fig. 1, Grand Theft Auto V (GTA) is a role-playing game with a city driving component and *realistic* traffic, making images closely resemble real driving scenes. In contrast, Synthia is a rendered dataset of a virtual city with *unrealistic* traffic simulation of dynamic objects like cars and people. Additionally, many Synthia images are not captured from a vehicle's perspective, making them less representative of real driving scenes.

Table 1 shows that our method significantly improves domain adaptation from GTA $\rightarrow$ Cityscapes, while Table 2 shows only marginal improvement for Synthia $\rightarrow$ Cityscapes. This result aligns with our expectations for both source datasets.

*Equal contribution

Figure 1. **Synthetic to Real Domain Adaptation**. GTA [10] consists of realistic scenes from a vehicle's perspective with accurate road traffic simulation, closely resembling real-world driving conditions. In contrast, Synthia [11] features unrealistic virtual-world images that differ significantly from natural driving scenes.

Table 1. **Synthetic to Real Domain Adaptation: GTA → Cityscapes Semantic Segmentation.** Tested on Cityscapes Val. We were unable to reproduce DAFormer's [5] results averaged over 3 random seeds. Therefore, we present results with seed=0 for both DAFormer [5] and our method (DAFormer with instance-level saliency guidance). Our method shows improvement in many individual categories for GTA. Although our focus is on real datasets rather than synthetic datasets, GTA *closely resembles natural driving scenes due to its realistic traffic simulation and vehicle-perspective imagery* (Figure 1). These *realistic* object-background appearances with background dominance contribute to the effectiveness of our method.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
CBST [20]	45.9	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
DACS [15]	52.1	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
CorDA [16]	56.6	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
ProDA [19]	57.5	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
DAFormer [5]	68.3	**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
DAFormer (seed=0)	66.9	92.6	58.9	89.3	54.2	42.7	49.4	57.0	55.8	89.2	49.8	89.5	**72.7**	41.7	92.0	62.0	**82.8**	**71.3**	56.5	62.9
+ Ours	**68.5**	92.9	60.0	**89.8**	**55.9**	**51.5**	49.0	**57.2**	**62.2**	89.6	**50.2**	91.5	71.9	**44.8**	**93.0**	**78.7**	79.8	63.6	**56.6**	**63.6**

Table 2. **Synthetic to Real Domain Adaptation: Synthia → Cityscapes Semantic Segmentation.** Tested on Cityscapes Val. We were unable to reproduce DAFormer's [5] results averaged over three random seeds. Therefore, we present results with seed=0 for both DAFormer [5] and our method (DAFormer with instance-level saliency guidance). Our improvement on Synthia → Cityscapes is marginal, which we attribute to the fact that *Synthia does not exhibit appearance akin to natural driving scenes* (Figure 1). These *unrealistic* object-background appearances make our method ineffective.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
CBST [20]	42.6	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
DACS [15]	48.3	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
CorDA [16]	55.0	**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
ProDA [19]	55.5	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
DAFormer [5]	**60.9**	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
DAFormer (seed=0)	59.2	87.8	47.6	87.6	43.7	5.8	49.0	48.0	53.1	82.3	-	71.8	71.7	46.0	87.5	-	49.3	-	52.7	**63.4**
+ Ours	59.4	92.7	57.7	87.1	**44.8**	**7.1**	46.2	46.9	53.6	80.6	-	68.2	72.5	42.5	**90.6**	-	41.6	-	**55.2**	63.3

Table 3. **Comparison with other Domain Adaptation Strategies: Computational Efficiency.** We use a single RTX 4090 GPU with a batch size of 1, maintaining the same training and inference image size of 512x1024 as specified in DAFormer [5]. We also include a comparison using a single A40 GPU, which has more memory, enabling us to experiment with an image size of 768x1536.

Our saliency-guided image warping imposes minimal memory overhead during training. Additionally, since we do not perform warping during test time, our method incurs no inference latency overhead.

Image Size	Method	Training Memory	Inference Time
512x1024	DAFormer	16.41 GB	195.0 ms
	DAFormer + Ours	**+ 0.04 GB**	**+ 0 ms**
	DAFormer + HRDA	+ 5.11 GB	+ 801.2 ms
768x1536	DAFormer	30.40 GB	261.3 ms
	DAFormer + Ours	**+ 0.04 GB**	**+ 0 ms**
	DAFormer + HRDA	+ 6.69 GB	+1074.0 ms

B. Comparison with other Domain Adaptation Strategies

B.1. Differences between HRDA/MIC and DAFormer

Domain adaptation extensions have been proposed for the DAFormer [5] training framework, such as HRDA [6] and MIC [7], both of which demonstrate strong performance. HRDA [6] introduces a multi-scale high-resolution crop training strategy combined with sliding window inference, whereas MIC [7] implements a masking consistency strategy for target domain images to enhance spatial context.

B.2. Computational Efficiency

Table 3 shows the computational efficiency comparison using the DAFormer image scale. We observe that incorporating HRDA into DAFormer significantly increases training and inference computational costs due to HRDA's train-time HR-cropping and test-time sliding window inference. In contrast, our inference costs *remain the same* as those of DAFormer [5], *as we do not warp during test time*. Our training memory consumption are only slightly higher, due to the lightweight design of our warping-related modules. As we noted in the main manuscript, no additional learned parameters are introduced. Moreover, the latency of the warping and unwarping operations is minimal, at 1.5 ms and 4.2 ms respectively.

B.3. Evaluation Methodology

Based on the above discussion, we apply HRDA's and MIC's strategies to DAFormer. However, for a fair comparison, we perform training and inference *following DAFormer's [5] training/testing image scales and evaluation paradigm*. An alternative comparison would involve training and evaluating on full-scale images to match HRDA, which deviates from original DAFormer [5] training and evaluation that uses half-scaled images. Unfortunately, we lack access to a GPU with high memory capacity, such as Tesla A100 with 80 GB of memory.

B.4. Results and Analysis

Cityscapes → DarkZurich/ACDC Semantic Segmentation: Table 4 and 5 present results for Cityscapes → DarkZurich and Cityscapes → ACDC semantic segmentation, respectively. MIC (HRDA) refers to MIC and HRDA training strategies added to DAFormer. We compare DAFormer and MIC (HRDA) with and without our instance-level saliency guided image warping. We observe smaller improvements with our method when combined with MIC (HRDA) due to HRDA's use of multi-scale cropping. See next paragraph for explanations.

Ablation on Cityscapes → ACDC Semantic Segmentation: We studied the interaction of our saliency-guided image warping with MIC [7] and HRDA [6] individually, as our method is orthogonal and plug-and-play. Table 6 shows that *(a)* HRDA did not improve DAFormer when trained at DAFormer's scale (1024×512) instead of full scale (2048×1024). *(b)* Our method improve performance for MIC and HRDA, but neither combination outperform DAFormer+Ours. We investigate and find that our limited improvement on HRDA is due to their use of 'detail crop'. *The HRDA detail crop focuses on small object regions, which our warping already oversamples. This results in redundant efforts and reduce overall effectiveness.* To verify this, we remove the detail crop from HRDA and observe that our method showed better improvement on this HRDA* variant (see Table 6).

Table 4. **Comparison with other Domain Adaptation Strategies: Cityscapes → DarkZurich Semantic Segmentation.** Tested on DarkZurich Val. Our method improves IoU scores in conjunction with both DAFormer and MIC (HRDA) strategies. We observe smaller improvements with MIC (HRDA) because both our warping and HRDA's detail-crop focus on small object regions, leading to redundant efforts.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
DAFormer	33.8	**92.0**	66.7	47.3	25.9	**50.8**	38.0	24.6	19.4	**62.2**	31.9	17.9	20.4	**28.4**	61.7	-	-	-	21.0	**34.3**
DAFormer + Ours	**37.1**	88.7	**70.9**	**60.1**	**42.1**	49.6	**39.8**	**47.9**	**21.5**	49.8	**35.7**	**25.4**	**23.8**	25.9	**66.8**	-	-	-	**24.3**	32.4
MIC (HRDA)	39.8	**78.8**	**13.0**	83.1	46.7	52.2	42.2	**44.5**	28.8	64.3	35.4	**82.8**	24.9	**33.8**	62.8	-	-	-	**25.9**	**37.3**
MIC (HRDA) + Ours	**40.1**	75.8	5.8	**83.3**	**55.7**	**54.8**	**44.3**	30.5	**38.0**	**66.3**	**42.7**	82.4	**39.4**	15.1	**69.3**	-	-	-	25.1	32.9

Table 5. **Comparison with other Domain Adaptation Strategies: Cityscapes → ACDC Semantic Segmentation.** Tested on ACDC Val. Our method improves IoU score in conjunction with both DAFormer and MIC (HRDA) strategies. We observe smaller improvements with MIC (HRDA) because both our warping and HRDA's detail-crop focus on small object regions, leading to redundant efforts.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
DAFormer	57.6	72.7	**57.5**	**80.1**	42.5	38.0	50.9	45.1	50.0	71.1	**38.5**	67.0	56.0	29.9	81.8	76.6	78.9	79.9	40.8	36.7
DAFormer + Ours	**61.8**	**82.9**	56.1	79.8	**44.6**	**40.3**	**52.7**	**60.8**	**52.5**	**72.0**	38.4	**78.0**	**56.6**	**30.5**	**84.9**	**80.2**	**86.9**	**86.4**	**44.5**	**45.8**
MIC (HRDA)	59.6	66.4	**55.8**	**84.8**	**54.6**	**41.3**	49.7	41.2	**53.6**	**78.3**	37.7	67.4	58.3	24.0	83.6	70.1	88.8	87.9	**40.7**	48.4
MIC (HRDA) + Ours	**61.6**	**80.6**	38.6	83.7	50.5	**41.3**	**50.5**	**56.7**	49.6	75.2	**39.4**	**83.9**	**58.9**	**32.0**	**86.0**	**75.4**	**91.8**	**88.0**	40.2	**48.5**

Table 6. **Comparison with other Domain Adaptation Strategies: Ablation on Cityscapes → ACDC Semantic Segmentation.** Tested on ACDC Val. Our method improves performance when combined with base DAFormer [5], MIC [7], HRDA [6], and HRDA without HR-detail crop (HRDA*). Our observations are – (a) Instance-level saliency guidance enhances MIC but does not exceed the performance of DAFormer with instance-level saliency guidance. (b) HRDA performs worse than DAFormer at DAFormer's training scales, indicating the necessity for full-resolution training for HRDA to perform well, consistent with findings in [6]. – (c) Removing the HR-detail crop (HRDA*) allows adding our warping method to achieve greater performance improvements. This is because both our warping and HRDA's detail crop focus on small object regions, resulting in redundant efforts.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
DAFormer	57.6	72.7	**57.5**	**80.1**	42.5	38.0	50.9	45.1	50.0	71.1	**38.5**	67.0	56.0	29.9	81.8	76.6	78.9	79.9	40.8	36.7
DAFormer + Ours	**61.8**	**82.9**	56.1	79.8	**44.6**	**40.3**	**52.7**	**60.8**	**52.5**	**72.0**	38.4	**78.0**	**56.6**	**30.5**	**84.9**	**80.2**	**86.9**	**86.4**	**44.5**	**45.8**
MIC (DAFormer)	58.8	61.1	59.5	73.2	**47.4**	**45.2**	51.4	44.7	48.2	**78.4**	38.1	51.4	60.4	**41.3**	84.3	78.5	84.3	78.9	**43.9**	**46.5**
MIC (DAFormer) + Ours	**60.6**	**72.8**	**62.9**	**73.4**	45.1	36.5	**52.9**	**49.0**	**49.9**	76.9	**39.8**	**65.5**	**60.6**	40.4	**85.2**	**80.9**	**90.5**	**87.0**	41.0	41.3
HRDA (DAFormer)	56.9	79.9	37.8	**81.1**	**45.6**	33.9	47.8	**47.3**	47.1	74.3	37.1	**84.0**	47.7	17.6	**84.2**	69.1	88.2	75.1	37.6	45.1
HRDA (DAFormer) + Ours	**57.7**	**85.6**	**48.2**	71.7	41.6	**39.4**	**50.8**	17.7	**47.8**	**75.2**	**37.9**	81.4	**56.3**	**25.0**	82.3	**73.4**	**88.8**	**82.3**	**45.3**	**46.7**
HRDA* (DAFormer)	58.3	68.5	59.4	82.8	**50.4**	**40.8**	50.3	42.4	44.4	**77.6**	38.0	69.4	**55.7**	27.2	83.2	77.4	78.2	79.0	34.9	48.4
HRDA* (DAFormer) + Ours	**62.1**	**89.7**	**61.1**	**83.9**	43.4	39.5	**52.7**	**43.1**	**45.0**	75.6	**38.7**	**86.1**	55.0	**28.0**	**84.9**	**81.1**	**88.5**	**86.0**	**44.9**	**53.4**

C. Supervised Setting

Supervised setting is not the focus of this work. Instance-level saliency guidance is effective in domain adaptation because its focus on objects significantly reduces the negative impact of backgrounds with large cross-domain variations. In supervised settings, where background variations are low within a single-domain, this focus on objects does not provide the same benefit.

In Section 3.2 of the main paper, we mentioned that source pre-training improves performance in the supervised setting on the source domain. Results are presented below.

Cityscapes Semantic Segmentation: We demonstrate improved semantic segmentation performance on Cityscapes by applying our method to the SegFormer model, which serves as the base architecture for the DAFormer [5] training strategy (see Table 7). Visual comparisons of our method versus SegFormer can be observed in Figure 2.

BDD100K Object Detection: We demonstrate improved object detection performance on BDD100K (Day) and BDD100 (Clear) when applying our method to Faster R-CNN, the base architecture of the 2PCNet [8] training strategy (see Table 8 and Table 9). Note that BDD100K (Day) includes images taken during the day in both clear and bad weather, while BDD100K (Clear) includes images taken in clear weather during both day and night.

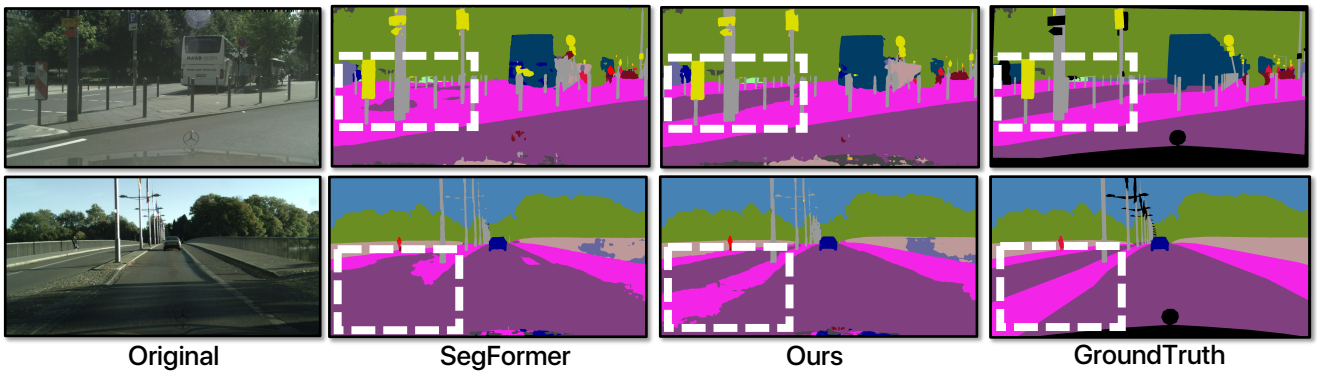

Figure 2. **Supervised Setting: Cityscapes Semantic Segmentation**. Our segmentation map more closely resembles the ground truth, indicating a more accurate understanding of objects and backgrounds in urban scenes. Notably, our method effectively distinguishes backgrounds such as sidewalks and roads, even in the presence of occlusions (top row) and shadows (bottom row).

Table 7. **Supervised Setting: Cityscapes Semantic Segmentation**. Tested on Cityscapes Val. Instance-level saliency guided image warping improves segmentation on the source domain by **+1.5 mIoU** (along with improvements on the target domain, shown in Tables 4 and 5 in the main manuscript), indicating better learned backbone features.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
SegFormer	75.3	98.0	83.4	91.8	59.6	59.6	57.5	64.0	74.4	91.9	62.8	94.6	77.6	56.0	93.7	81.6	81.4	70.1	59.2	73.3
+ Ours w/ Sta. Prior	76.1	97.9	83.6	91.9	58.0	58.0	57.7	63.3	73.9	91.7	64.4	94.4	77.7	**57.4**	93.8	81.0	**87.7**	80.3	60.7	72.8
+ Ours w/ Geo. Prior	76.5	98.0	84.3	**92.2**	**61.4**	57.6	59.3	64.7	74.0	91.9	65.3	94.7	78.2	55.3	93.9	**83.1**	86.5	81.1	60.1	72.1
+ Ours w/ Inst.	**76.8**	**98.1**	**84.8**	**92.2**	59.9	**58.3**	**59.6**	**65.1**	**75.4**	**92.3**	**66.2**	**94.8**	**78.2**	55.3	**94.2**	82.0	85.7	**81.3**	**61.8**	**74.4**

Table 8. **Supervised Setting: BDD100K (Day) Object Detection.** Tested on BDD100K Day Val, which includes images captured in both good and bad weather. As shown by mAP50 (overall and per category), our saliency guided image warping improves detection performance in the source domain, and our instance-level saliency guidance is competitive compared with other saliency priors.

Method	mAP	**mAP50**	mAP75	mAPs	mAPm	mAPl	person	rider	car	truck	bus	motor cycle	bike	traffic light	traffic sign
FRCNN	30.1	56.4	28.1	13.9	37.6	51.0	64.0	50.7	80.3	62.5	62.9	45.3	49.7	66.7	69.8
+ Ours w/ Sta. Prior	30.9	57.1	**28.6**	14.4	**38.9**	**53.4**	65.7	53.1	**80.9**	**62.7**	63.0	**48.8**	50.8	**69.2**	71.2
+ Ours w/ Geo. Prior	**31.1**	**57.9**	28.3	**14.5**	38.5	52.7	66.3	**53.6**	80.7	62.5	62.8	48.1	**52.9**	68.4	**71.4**
+ Ours w/ Inst.	30.7	57.2	27.9	**14.5**	38.4	52.8	**66.4**	53.3	80.8	62.4	**63.7**	47.7	51.7	68.6	71.1

Table 9. **Supervised Setting: BDD100K (Clear) Object Detection.** Tested on BDD100K Clear Val, which includes both day and night images. As shown by mAP50 (overall and per category), our saliency guided image warping improves detection performance in the source domain, and our instance-level saliency guidance is competitive compared with other saliency priors.

Method	mAP	**mAP50**	mAP75	mAPs	mAPm	mAPl	person	rider	car	truck	bus	motor cycle	bike	traffic light	traffic sign
FRCNN	25.4	49.6	22.5	12.2	30.3	44.2	59.3	38.8	76.5	53.2	54.7	43.1	45.6	56.4	68.7
+ Ours w/ Sta. Prior	**26.0**	50.2	**22.9**	11.9	**31.2**	44.5	59.2	38.8	**76.7**	**54.5**	55.7	**45.5**	46.0	56.6	69.2
+ Ours w/ Geo. Prior	25.9	**50.3**	22.8	12.0	31.0	**44.8**	59.4	**41.1**	**76.7**	53.7	56.1	42.7	**46.9**	**56.7**	**69.3**
+ Ours w/ Inst.	25.9	50.1	22.8	**12.2**	31.0	43.7	**59.7**	37.7	76.6	54.4	**56.9**	43.7	46.0	56.5	69.1

D. Additional Domain Adaptations

Cityscapes → Foggy Cityscapes: Table 10 shows the result for domain adaptative object detection from Cityscapes → Foggy Cityscapes. Our improvement is small on the Foggy Cityscapes dataset. This is because *(a)* the baseline is already strong when dealing with easy synthetic fog *(b)* there is little cross-domain background variation, leaving minimal room for improvement with our warping approach that oversamples objects to reduce the negative impacts of cross-domain large background variations.

Table 10. **Additional Domain Adaptations: Cityscapes → Foggy Cityscapes.**. We observe small improvements because – (a) The *synthetic* fog introduced in Foggy Cityscapes does not accurately mimic real fog, thereby posing less of a challenge to segmentation algorithms. Consequently, the baseline model already attains high scores, making it difficult for our warping method to yield substantial improvements. – (b) Since Foggy Cityscapes adds a fixed amount of fog to each image in Cityscapes, there is minimal cross-domain background variations. In this situation, our warping strategy that oversamples objects to mitigate negative background impacts from large cross-domain background variations is less effective.

Method	**mIoU**	road	side walk	building	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor cycle	bike
DAFormer	74.7	97.8	83.0	88.7	**60.4**	59.4	57.4	**60.7**	72.8	89.2	**64.9**	81.2	76.6	54.4	**93.2**	75.9	86.9	80.7	62.8	73.8
+ Ours w/ Inst.	**75.5**	**98.0**	**84.2**	**89.7**	59.7	**61.9**	**57.7**	60.4	**73.3**	**89.4**	62.2	**82.4**	**76.7**	**56.3**	92.8	**82.2**	**88.0**	**82.0**	**63.5**	**73.9**

E. Additional Analysis

Grad-CAM Visualizations of the Learned Model: In Section 5.2 of the main paper, we claimed that the learned backbone features obtained through our training method generalize better. To support this claim, we use Grad-CAM [14] visualizations to showcase ResNet features from detectors trained on BDD100K [18], as shown in Figure 3. We observe that *(a)* Heatmaps for models trained with 2PCNet show that with our warping, there is a strong focus on salient objects with minimal distracting activation, while without our warping, the focus is dispersed across the image, indicating a lack of precision. *(b)* The choice of saliency guidance matters: our instance-level saliency guidance ensures a strong focus on salient objects, whereas warping guided by Geometric Prior [3] results in less emphasis on target objects (e.g., cars) and an unnecessary focus on the background.

Grad-CAM Visualizations for Multiple Objects: Using the same backbone layers, Figure 4 shows that Grad-CAM contributions from background pixels are smaller compared to those from foreground object pixels when our method is used during training. This demonstrates improved focus on foreground objects over the background when trained with our instance-level saliency guidance.

Per-Pixel Accuracy Difference Visualizations: Figure 5 shows per-pixel accuracy difference visualizations. For both Cityscapes [2] and ACDC [13], a noticeable predominance of red over blue is observed, indicating a clear advantage of our method over DAFormer [5] for semantic segmentation.

BDD100K Clear $\rightarrow$ DENSE Foggy Object Detection: Qualitative comparisons are shown in Figure 6. Our method demonstrates superior object detection under real foggy conditions by accurately identifying objects like streetlights and vehicles. In contrast, 2PCNet [8] misidentifies windows as pedestrians, a mistake our approach avoids.

BDD100K (Clear $\rightarrow$ Rainy) Object Detection: Qualitative comparisons are shown in Figure 7. Our method demonstrates superior object detection under rainy conditions by accurately identifying vehicles and minimizing false positives, such as misidentified pedestrians and cars in the 2PCNet [8] predictions.

Cityscapes $\rightarrow$ DarkZurich Semantic Segmentation: Qualitative comparisons are shown in Figure 8. The proposed method produces segmentation outputs that more closely align with the ground truth, particularly in predicting road boundaries and consistently identifying urban elements such as sidewalks, terrain, and traffic signs. This indicates that our method has superior domain adaptation capabilities in challenging low-light conditions.

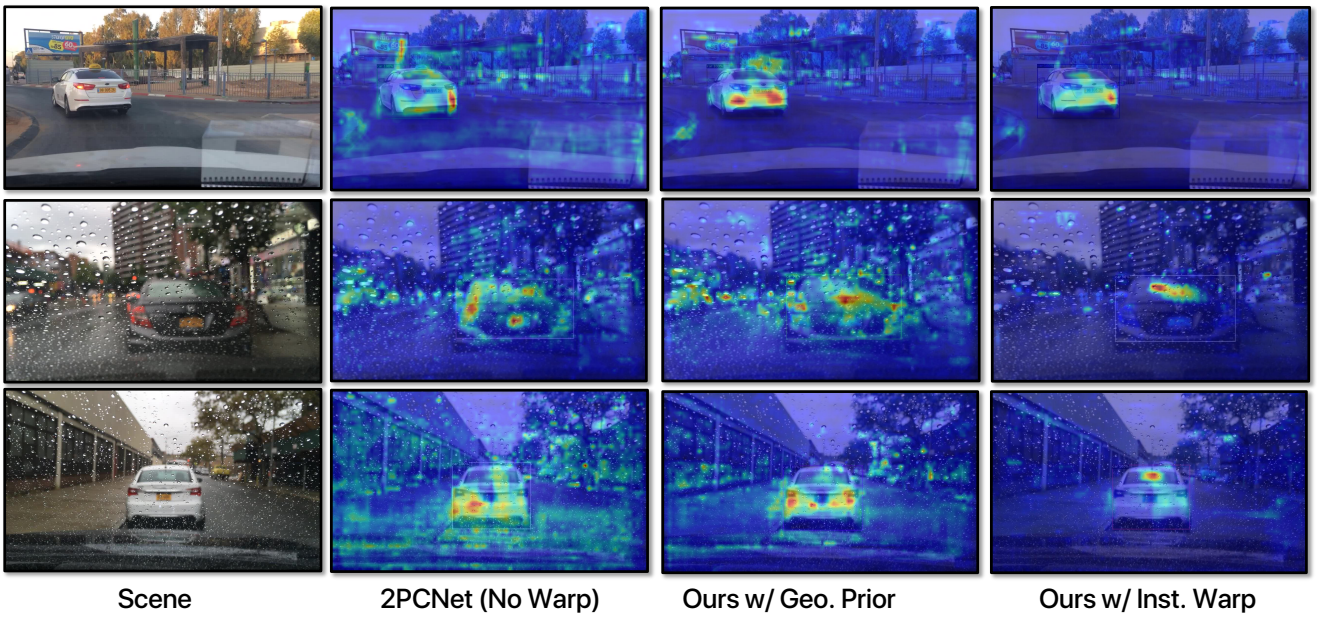

Figure 3. **Additional Analysis: Grad-CAM [14] Visualization of the Learned Model**. We visualize the last layer feature of the learned ResNet-50 backbone. Grad-CAM visualization shows that the model trained with our method demonstrate a higher focus on salient objects, indicating better-learned features and improved scene comprehension.

Figure 4. **Additional Analysis: Grad-CAM [14] Visualizations for Multiple Objects.** We observe that learnt features show smaller Gradcam visualization contributions from background pixels compared to foreground object pixels when trained with instance-level warping. This is true for individual object instances, in this case, for Vehicle 1 and Vehicle 2. This suggests that our instance-level warping enhances the model's focus on foreground object pixels over background pixels.

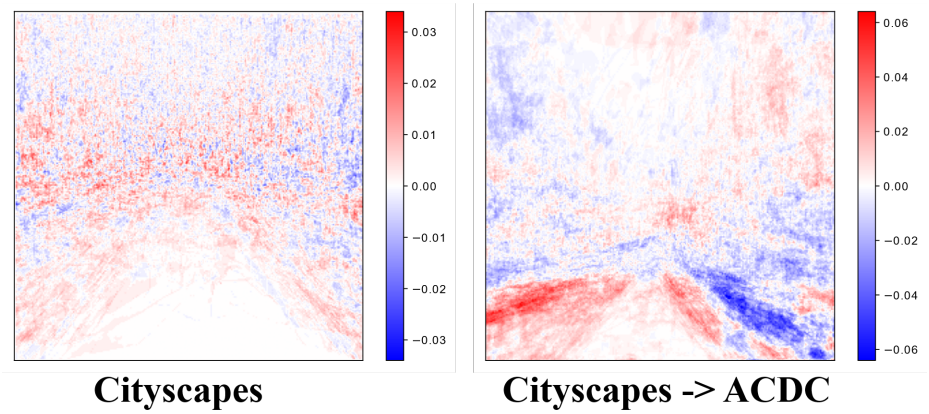

Figure 5. **Additional Analysis: Per-Pixel Accuracy Difference Visualization**. The visualization is done between baseline and ours. Red indicates our method is better, whereas blue means the baseline is better. To improve the quality of visualization, we omit the comparison for 'sky' and reshape the semantic segmentation maps to 256 × 256. We notice that our method outperforms the baseline across most pixels, except for the right-hand side sidewalk pixels in the ACDC dataset (blue band) due to a significant disparity in width of the sidewalk (a background element) compared to the Cityscapes dataset.

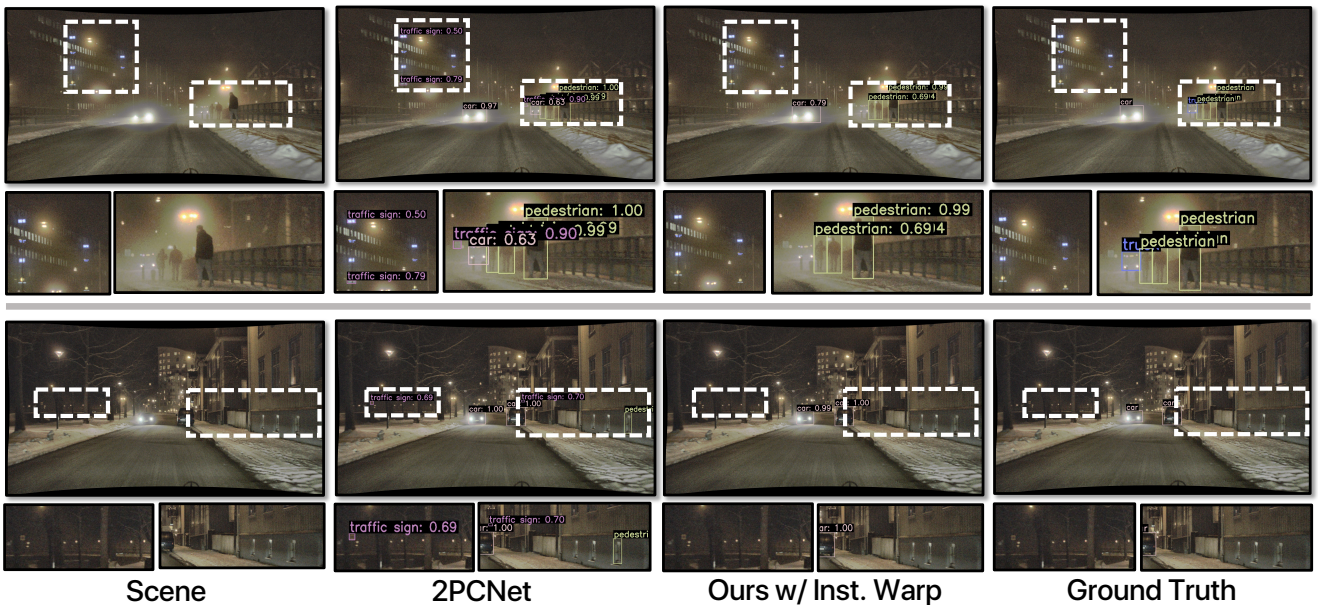

Figure 6. **Additional Analysis: BDD100K Clear → DENSE Foggy Object Detection.** Our method demonstrates superior object detection under foggy conditions. It accurately identifies streetlights and vehicles, which 2PCNet mislabels as traffic signs. Additionally, our method correctly ignores windows, which 2PCNet misclassifies as pedestrians.

Figure 7. **Additional Analysis: BDD100K (Clear → Rainy) Object Detection.** Our method demonstrates superior object detection under rainy conditions by accurately identifying vehicles and minimizing false positives, such as misidentified pedestrians and cars evident in 2PCNet [8] predictions.

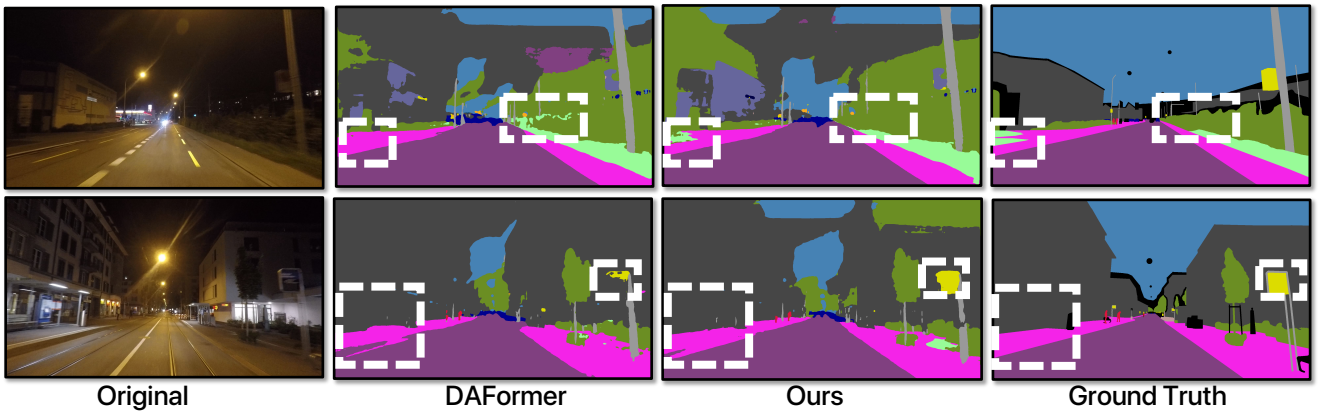

Figure 8. **Additional Analysis: Cityscapes → DarkZurich Semantic Segmentation.** Our method exhibits superior semantic segmentation in challenging low-light conditions. It produces segmentation outputs that closely align with the ground truth, particularly in accurately predicting road boundaries and consistently identifying urban elements such as sidewalks, terrain, and traffic signs.

F. Experimental Methodology

Object Detection: Following the recent and popular domain adaptation strategy 2PCNet [8], we use Faster R-CNN [9] with ResNet-50 [4], adhering to their training hyperparameters and protocols. While 2PCNet [8] focuses solely on Day-to-Night adaptation, our method addresses various adaptation scenarios. For scenarios other than Day-to-Night adaptation, we exclude the NightAug augmentation proposed by 2CPNet for both their method and ours.

Semantic Segmentation: We follow DAFormer [5] by employing the same SegFormer [17] head and MiT-B5 [17] backbone, adhereing to their training hyperparameters, protocols, and seed for fair comparison. While Sim2Real Gap result presented in DAFormer [5] are not our focus, relevant results and discussion can be found in Supp. A.

Datasets: We use BDD100K [18], Cityscapes [2], DENSE [1], ACDC [13] & DarkZurich [12]. A brief description is given below.

BDD100K [18] features 100,000 images with a resolution of 1280x720 for object detection and segmentation, covering various weather conditions and times of day, and includes annotations for 10 categories.

Cityscapes [2] provides 5,000 images of urban road scenes at a resolution of 2048x1024 in clear weather for semantic segmentation, with annotations for 19 categories.

DENSE [1] provides 12,997 images at a resolution of 1920x1024, capturing diverse weather conditions such as heavy fog and heavy snow.

ACDC [13] is designed for adverse conditions such as fog and snow, including 1,600 images at a resolution of 2048x1024 for segmentation across 19 categories.

Dark Zurich [12] is tailored for low-light conditions, offering 2,416 unlabeled nighttime images and 151 labeled twilight images for segmentation, all at a resolution of 1920x1080, with a focus on urban settings.

G. Additional Technical Details

Ground Truth Segmentation to Boxes: Instance-level saliency guided image warping requires bounding boxes, which are not provided in some semantic segmentation benchmarks like GTA [10]. To address this, we generate 'from-seg' bounding boxes from ground truth semantic segmentation maps. Specifically, we first identify connected components representing individual instances of foreground categories, including traffic lights, traffic signs, persons, riders, cars, trucks, buses, trains, motorcycles, and bikes. For each 'from-seg' instance, we then compute the bounding boxes by finding the minimum enclosing axis-aligned rectangle. These 'from-seg' bounding boxes are finally used for instance-level saliency guidance in the same way as ground truth bounding boxes.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 2020. 13

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 9, 13

[3] Anurag Ghosh, N Dinesh Reddy, Christoph Mertz, and Srinivasa G Narasimhan. Learned two-plane perspective prior based image resampling for efficient object detection. In *CVPR*, 2023. 9

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 13

[5] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 2, 3, 4, 5, 9, 13

[6] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 3, 4

[7] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 3, 4

[8] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *CVPR*, 2023. 5, 9, 12, 13

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 13

[10] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 1, 2, 14

[11] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 1, 2

[12] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019. 13

[13] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 9, 13

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 9, 10

[15] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 2

[16] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, 2021. 2

[17] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 13

[18] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 9, 13

[19] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 2021. 2

[20] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2