Hybrid-TTA: Continual Test-time Adaptation via Dynamic Domain Shift Detection

Supplementary Material

1. Implementation Details

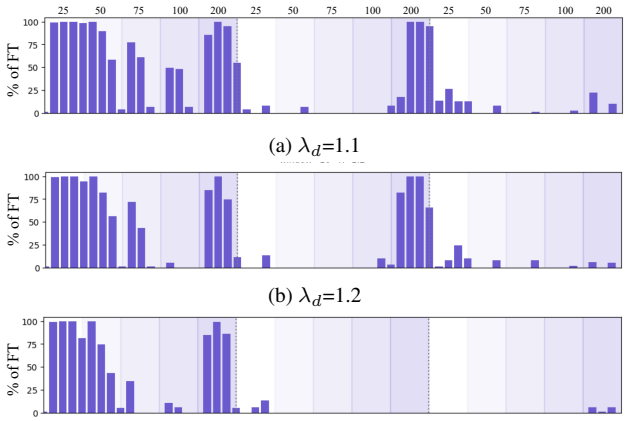
We report the implementation details used to train in the described method, including network architectures, and hyperparameters.

Model Architecture For semantic segmentation, we use SegFormer MiT-B5 for feature extractor f_θ and segmentation decoder h_ϕ. For reconstruction, SimMIM [43] is used for the implementation of reconstruction decoder g_ψ. To implement Adapter Tuning, we utilize AdaptMLP from [3] into each transformer layers to implement efficient tuning, where Full Tuning updates all parameters including the adapter, and Adapter Tuning updates only the adapter parameters (which occupies 0.1% of the entire parameters).

Hyperparameters For test-time adaptation, we set the batch size to 1, and used Adam optimizer with learning rate of $6 \times 10^{-5}/8$, in reference of [25, 41]. For DDSD implementation, we set the initial DDSD threshold τ to be 0, and default α_l to be 0.999, following [37]. For MIMA implementation, we masked the images with masking ratio 0.6 and mask patch size 32, as in [43]. Unlike recent CTTA studies [25, 41, 45] that used multi-scale input with flipping as the test-time Augmentation, we did not use any augmentation strategy for our main results. λ_d=1.1 and λ_r=0.3 were heuristically selected based on a small subset of the first target domain and kept fixed across all experiments to avoid test-time overfitting. The same hyperparameters are used for both benchmarks.

2. Ablation Study

Hyperparameter λ_d We conducted ablation experiments to investigate the role the hyperparameter λ_d in DDSD works within the system, using OnDA benchmark [29]. Only the segmentation loss L_{seg} has engaged in this ablation study to exclude the influence of MIMA. We performed cyclic adaptation over 5 rain intensities (25mm-200mm) 3 times, to understand how DDSD adapts to a completely new domains and how it reacts after some initial adaptations. Since 25mm is closer to the source domain and 200mm farther, we consider 200mm to be harder to adapt to, requiring more adaptation than 25mm. The domain changes are visualized by changing background colors, according to the changing rain intensities written at the top. The height of each bar represents the ratio of Full Tuning (FT) being activated in every 125 timesteps, *e.g.,* if FT is activated 100 times out of 125, the ratio is 80%.



(a) λ_d=1.1

(b) λ_d=1.2

(c) λ_d=1.3

Figure 7. **Ablation on hyperparameter λ_d.** Percentage of FT usage in OnDA benchmark. x-axis represents time and y-axis represents FT usage ratio. 5 domains are cyclically repeated 3 times.

Fig 7 illustrates how DDSD evolves over time. After aggressive adaptation in the first round, triggered by the initial encounter with the target domain, DDSD begins to favour AT over FT, leading to a reduced FT ratio. This change occurs because the target domain becomes relatively familiar to both the student and teacher models. However, in 200mm domain during the first round, the FT ratio increases, as 200mm is significantly more distant from any other domains and thus exhibits lower temporal correlation. This phenomenon consistently holds across all experiments, regardless of the size of λ_d value.

The effect of hyperparameter λ_d is particularly evident in 100mm domain during the first round. In Fig. 7a, the FT ratio is nearly 50%, while in Fig. 7b and Fig. 7c, the ratio drops sharply to below 10%. Similarly, in 200mm domain during the second round, FT is not activated in Fig. 7c, unlike the other two graphs.

From these observations, we can conclude that as λ_d increases, DDSD becomes less sensitive to domain shifts. This is due to the role of λ_d a sensitivity hyperparameter (described in Sec. 4.1), which acts as a scaling factor for the dynamic threshold τ_t. As the dynamic threshold increases, a larger loss is required to activate FT, and vice versa.

Hyperparameter λ_r In Tab. 6, we conducted ablation studies with respect to λ_r in (6), which determines the in-

Table 5. Ablation Study in Cityscapes-to-ACDC benchmark for **20 rounds.** Mean is the average score of mIoU.

Test Condition	FT	DDSD	MIMA	1 F	1 N	1 R	1 S	10 F	10 N	10 R	10 S	20 F	20 N	20 R	20 S	Mean ↑ 10R	Mean ↑ 20R
(a)	✓			70.6	42.7	62.7	61.7	69.3	42.7	61.2	59.7	65.5	39.3	55.9	58.5	59.3	55.9
(b)	✓		✓	70.3	44.5	65.1	63.3	68.9	49.2	64.9	61.5	66.2	48.2	62.0	57.5	61.9	59.6
(c)		✓		70.9	43.2	64.0	60.7	70.5	43.0	64.3	60.5	69.6	43.6	62.5	59.1	59.8	58.9
(d) - ours		✓	✓	70.3	44.5	65.1	63.2	69.9	49.5	66.5	63.0	68.8	50.1	63.9	59.9	62.2	61.5

Table 6. **Ablation on hyperparameter** λ_r Performance in Cityscapes-to-ACDC benchmark.

	λ_r	**Performance (mIoU)**
(a)	0.0	61.3
(b)	**0.3**	**61.9**
(c)	0.5	61.3
(d)	1.0	60.6

fluence of the reconstruction loss $\mathcal{L}^{rec}_{\theta,\psi}$. To exclude the influence of DDSD, we only utilized Full Tuning.

In (a), λ_r is set to 0, meaning that the reconstruction loss does not participate in test-time Adaptation process. This setup is similar to MIC [17], where the model is adapted solely through the consistency loss between student and teacher prediction maps. Since some important visual information is masked out from the target image, (a) yields a suboptimal performance of 61.3%. In (b), λ_r is set to 0.3, achieving the best performance. However, as λ_r increases in (c) and (d), performance declines. This degradation occurs because the influence of the segmentation loss diminishes as reconstruction loss dominates the adaptation process with big λ_r. Consequently, adaptation towards segmentation, which is our primary task, becomes weaker.

Long-term Adaptation Although 0.5%p improvement of DDSD in Tab. 3 may appear minor, we believe its core strength–stability–has been underestimated within the scope of the original 10-round experiment. Fig. 5 clearly illustrates that DDSD maintains stable performance, whereas Full Tuning (FT) exhibits a sharp performance decline after the 5$^{\text{th}}$ round, indicating that DDSD's advantage will grow over time. To further support this observation, we extend the ablation study reported in Tab. 3 from 10 rounds to 20 rounds. As shown in Tab. 5, (c) DDSD consistently outperforms FT at both the 10$^{\text{th}}$ and 20$^{\text{th}}$ rounds, with an widening performance gap (0.5→**3.0** between (a) FT and (c) DDSD, 0.3→**1.9** between (b) FT+MIMA and (d) DDSD+MIMA). These results strongly confirm the superior long-term stability of DDSD.

3. Analysis

Source Domain Forgetting Tab. 7 presents the performance on Cityscapes dataset after each round of adaptation on OnDA benchmark. Since Cityscapes serves as the source dataset for adaptation, we aim to assess the severity of catastrophic forgetting after each round of adaptation. 'Mean' represents the mIoU performance on Cityscapes, while 'Target Mean' refers to the performance on synthetic rain dataset, provided for comparison.

Table 7. Performance comparison of variants of proposed method on **Source Dataset** (CityScapes). *ours* refers to Hybrid-TTA, our main result, *Source* refers to SegFormer MiT-B5 with no adaptation.

Test	1	2	3	4	5	Mean	Target Mean
(a) Source	78.0	78.0	78.0	78.0	78.0	78.0	62.4
(b) AT	77.7	77.4	77.2	77.3	77.0	**77.3**	60.5
(c) DDSD	76.7	76.2	75.6	75.4	75.2	75.8	64.2
(d) DDSD+MIMA	77.0	76.0	75.3	75.0	74.7	75.5	**66.7**
(e) FT	76.5	75.6	75.2	74.9	74.7	75.4	65.0
(f) AT+MIMA	77.3	75.9	74.8	74.0	73.1	75.0	65.3
(g) FT+MIMA	76.8	75.7	74.8	74.2	73.7	75.0	66.3

(a) Source, SegFormer MiT-B5 without any adaptation, serves as the upper bound, achieving 78% on the source dataset, with moderate target performance of 62.4%. (b) AT, which updates only the adapter parameters, achieves the best source performance of 77.3%, but its target performance (60.5%) is lower than Source. This is because AT effectively preserves source knowledge and prevents catastrophic forgetting but has limited adaptability. However, this limitation is significantly mitigated by incorporating MIMA into AT, as (f) AT+MIMA shows 65.3% of target performance at the cost of slight reduction in source performance.

Here, we can assume that MIMA encourages the model to lose the source knowledge and collect target knowledge, as a similar pattern is observed with (e) FT and (g) FT+MIMA, where (g) exhibits lower source performance but higher target performance compared to FT. As seen from these findings, losing source knowledge is not necessarily disadvantageous, because the model learns as much as it forgets.

Nevertheless, excessive loss of source knowledge can degrade target performance. Notably, (d) DDSD+MIMA (ours) achieves an impressive 66.7% mIoU on the target dataset, with a gain of $+4.3\%p$ compared to Source, while substantially preserving source performance ($-2.5\%p$).

Table 8. Performance comparison in **OASIS benchmark**. GTA [30] as the source dataset, and ACDC [32] as the test dataset.

Test	1				2				3				
Condition	F	N	R	S	F	N	R	S	F	N	R	S	Mean ↑
Source	43.4	19.6	41.3	38.1	43.4	19.6	41.3	38.1	43.4	19.6	41.3	38.1	35.6
TENT [40]	43.8	19.8	42.1	38.7	43.9	18.8	39.9	35.9	43.0	17.1	37.4	33.5	34.5
SVDP [45]	**46.9**	**25.2**	45.6	41.8	48.0	25.0	43.8	39.6	45.1	22.5	42.6	40.1	38.8
C-MAE [26]	46.1	20.0	43.2	38.9	45.8	18.9	42.9	39.1	46.2	20.5	43.2	38.1	36.9
Ours	45.1	23.0	**46.6**	**42.5**	**48.9**	**26.9**	**48.2**	**43.3**	**49.2**	**28.8**	**48.0**	**43.3**	**41.2**

Table 9. Performance Comparison on **Cityscapes-to-ACDC benchmark** over 3 rounds.

Test	1				2				3				
Condition	F	N	R	S	F	N	R	S	F	N	R	S	Mean↑
CoTTA [41]	70.9	41.2	62.4	59.7	70.9	41.1	62.6	59.7	70.9	41.0	62.7	59.7	58.6
VDP [10]	70.5	41.1	62.1	59.5	70.4	41.1	62.2	59.4	70.4	41.0	62.2	59.4	58.2
BeCoTTA [20]	**72.3**	42.0	63.5	60.1	**72.4**	41.9	63.5	60.2	**72.3**	41.9	63.6	60.2	59.5
C-MAE [26]	71.9	**44.6**	**67.4**	**63.2**	71.7	44.9	66.5	63.1	**72.3**	45.4	**67.1**	63.1	61.8
Ours	70.3	44.5	65.1	**63.2**	71.8	**48.2**	**67.1**	**63.7**	71.2	**49.3**	**67.1**	**63.3**	**62.2**

GTA-to-ACDC benchmark We now dive deeper in Hybrid-TTA performance on the OASIS [39] benchmark protocol, as detailed in Tab. 8. The OASIS protocol involves training models on a synthetic source dataset (GTA [30]), tuning hyperparameters on a synthetic validation dataset (SYNTHIA [31]), and finally evaluating model performance on a real-world test dataset (ACDC [32]). It is widely acknowledged that the OASIS benchmark presents greater challenge than the Cityscapes-to-ACDC benchmark due to a significantly larger domain gap between source and target datasets.

Our base model was trained on GTA following the training protocol described in [39]. We select the best hyperparameters in SYNTHIA dataset as follows: λ_d=1.2, λ_r=0.2. Finally, we present CTTA results on ACDC dataset over 3 rounds, where Ours outperforms SoTA [26, 45], demonstrating our robustness under various environments.

Although using FT under large domain shifts may seem counter-intuitive, severe shifts require greater adaptability than what efficient tuning can offer (*e.g.*, TENT in Tab. 8), at the cost of stability and with the forgetting risks associated with FT. To manage this trade-off and achieve balance, DDSD selectively triggers FT under significant distribution shifts, while avoiding unnecessary updates in stable regimes.

Performance Comparison with Parameter-efficient Fine-tuning Methods Tab. 9 provides a detailed comparative study of various CTTA strategies on the Cityscapes-to-ACDC benchmark, extending the results discussed in 5.2, but evaluated over 3 adaptation rounds to highlight early-stage adaptation behaviors and performances.

VDP [10], a pioneering work that introduces a prompt-based Parameter-efficient Fine-tuning (PEFT) strategy for CTTA, achieves a relatively disappointing result of 58.2, notably underperforming even the baseline CoTTA. This suggests that prompt-based PEFT adaptation strategies might struggle in scenarios involving significant domain shifts, such as from CityScapes to ACDC.

BeCoTTA [20], which employs a Mixture-of-Experts (MoE) based Parameter-efficient Fine-Tuning CTTA strategy, demonstrates slightly better average mIoU (59.5) compared to VDP (59.4).

Performance Comparison with Continual-MAE Tab. 9 also provides performance of Continual-MAE [26], another pioneering strategy based on MIM, which attains a considerably stronger performance of 61.8. While our method employing MIMA outperforms C-MAE, our method also differs significantly in methodology. Specifically, C-MAE reconstructs HOG features to emphasize geometric changes (*e.g.*, shape), promoting the learning of domain-invariant features. On the other hand, MIMA reconstructs RGB values, not only enhancing robust feature extraction, but also capturing domain-specific low-level cues (*e.g.*, color, texture, brightness). This enables the model to better capture cross-domain feature discrepancies, allowing DDSD to detect domain shifts more effectively.

While MIM has been previously explored in TTA as a domain-agnostic regularizer [26] or pseudo-label generator [28], our work departs from these conventional usages in both design and intent. The integration between MIMA and DDSD is not merely synergistic but functionally complementary: MIMA enhances sensitivity to domain discrepancies while improving robustness on familiar data, and DDSD leverages this sensitivity to selectively detect domain shifts. We believe this task-driven coupling of MIM and domain shift detection in CTTA is novel, offering a new perspective on how reconstruction-based signals can be actively utilized beyond pretraining.

4. Performance Comparison (Full scores)

We provide the entire performance results of segmentation CTTA experiments in Tab. 10 and Tab. 11. Our proposed method, Hybrid-TTA, achieves 0.6%p mIoU improvement over the previous state-of-the-art method on Cityscapes-to-ACDC benchmark. Moreover, it outperforms other CTTA methods on OnDA benchmark by 0.9%p. Notably, Hybrid-TTA also achieves more than 20 times higher FPS than any other CTTA method with comparable mIoU performance, including CoTTA and SVDP (See Fig. 1 and Tab. 4), offering a robust solution for real-world online continual adaptation challenges.

5. Qualitative Results

Fig. 8 is qualitative comparison of segmentation results on Fog, Night, Rain and Snow domains for Cityscapes-to-ACDC benchmark. Hybrid-TTA (column 4) is showing remarkable segmentation results compared to other methods including CoTTA [41] and SVDP [45], notably SVDP being currently the state-of-the-art method in semantic segmentation CTTA.

In Night domain (row 3-4), Hybrid-TTA shows outstanding performance as it is demonstrated in the main paper. It is noticeable in Hybrid-TTA (row 3-4, column 4), particularly in distinguishing the sky from vegetation (green) and buildings (gray), where lighting conditions are poor. This is a significant achievement, given that other methods often misclassify these elements due to the altered appearance of objects at night.

For Rain domain (row 5-6), Hybrid-TTA excels in segmenting fine details such as fence (beige) and accurately identifying cars (deep blue) and sidewalk (pink) from road (purple), which other methods often confuse with vegetation (green) or terrain (light green). This highlights the model's ability to maintain clear boundaries and accurate object classification under adverse weather conditions.

In case of Snow domain (row 7-8), Hybrid-TTA effectively delineates sidewalk (pink) and other urban features, producing sharper segmentation maps compared to CoTTA and SVDP, which often blur these boundaries.

Overall, we can observe that Hybrid-TTA not only maintains robustness across diverse environmental conditions but also mitigate common segmentation issues, such as dirty segmentation maps observed in CoTTA (row 3, column 2) and SVDP (row 6, column 3). This robustness is particularly evident in the Night domain, which is typically considered the most challenging due to its low contrast and ambiguous object boundaries. These qualitative results underscore the effectiveness of Hybrid-TTA in real-world scenarios, where adaptability and precision are crucial.

Table 10. Performance comparison of Segmentation CTTA baselines on **Cityscapes-to-ACDC benchmark** over 10 cyclic rounds. Mean is the average score of mIoU.

Test	1				2				3				4				5				
Condition	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	Cont.
Source	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	-
CoTTA [41]	70.9	41.2	62.4	59.7	70.9	41.1	62.6	59.7	70.9	41.0	62.7	59.7	70.9	41.0	62.7	59.7	70.9	41.0	62.8	59.7	-
SVDP [45]	71.0	43.2	64.8	62.0	71.9	44.3	66.1	62.5	72.2	44.1	66.6	62.5	71.8	43.9	66.6	62.3	72.1	43.7	66.7	62.5	-
ViDA [25]	71.6	43.2	66.0	63.4	73.2	44.5	67.0	63.9	73.2	44.6	67.2	64.2	70.9	44.0	66.0	63.2	72.0	43.7	66.3	63.1	-
Hybrid-TTA	70.3	44.5	65.1	63.2	71.8	48.2	67.1	63.7	71.2	49.3	67.1	63.3	70.1	49.3	66.9	63.1	69.5	49.2	66.6	63.0	-

Test	6				7				8				9				10				All↑
Condition	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	**Fog**	**Night**	**Rain**	**Snow**	Mean
Source	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	69.1	40.3	59.7	57.8	56.9
CoTTA [41]	70.9	41.0	62.8	59.7	70.9	41.1	62.6	59.7	70.9	41.1	62.6	59.7	70.8	41.1	62.6	59.7	70.8	41.1	62.6	59.7	58.6
SVDP [45]	71.9	43.9	66.8	62.4	71.9	43.6	66.5	62.3	71.6	43.3	66.6	62.1	71.7	43.7	66.5	61.5	71.6	43.5	66.6	62.0	61.0
ViDA [25]	72.2	44.0	66.6	62.9	72.3	44.8	66.5	62.9	72.1	45.1	66.2	62.9	71.9	45.3	66.3	62.9	72.2	45.2	66.5	62.9	61.6
Hybrid-TTA	69.5	49.3	66.7	63.0	69.6	49.4	66.7	63.0	69.7	49.5	66.7	63.0	69.7	49.4	66.6	63.1	69.9	49.5	66.5	63.0	**62.2**

Table 11. Performance comparison of Segmentation CTTA baselines on **OnDA benchmark.** over 5 cyclic rounds. Mean is the average score of mIoU.

Test	1					2—					3					4					5					All↑
Intensity (mm)	**25**	**50**	**75**	**100**	**200**	**25**	**50**	**75**	**100**	**200**	**25**	**50**	**75**	**100**	**200**	**25**	**50**	**75**	**100**	**200**	**25**	**50**	**75**	**100**	**200**	Mean
Source	67.7	65.6	62.9	58.1	45.3	67.7	65.6	62.9	58.1	45.3	67.7	65.6	62.9	58.1	45.3	67.7	65.6	62.9	58.1	45.3	67.7	65.6	62.9	58.1	45.3	59.9
TENT-continual [40]	66.8	64.8	62.3	58.1	45.6	65.8	63.3	60.7	56.5	44.5	64.2	61.3	58.6	54.6	43.1	62.6	59.5	56.7	53.0	42.0	61.1	58.1	55.4	51.8	41.2	56.5
CoTTA [41]	68.9	67.6	66.8	65.6	59.2	69.2	68.1	67.2	66.0	60.3	68.7	67.5	66.6	65.6	60.1	68.6	67.5	66.6	65.8	60.3	68.6	67.6	66.7	65.7	60.7	65.8
SVDP [45]	69.2	68.1	66.9	65.5	61.7	67.7	67.2	66.6	64.7	61.2	67.5	66.8	66.1	65.3	62.6	67.7	67.1	66.2	65.4	62.2	67.4	66.6	66.1	65.4	62.1	65.7
Hybrid-TTA	68.2	67.7	67.4	66.8	62.9	68.7	68.2	67.5	66.8	64.4	68.1	67.7	67.1	66.5	64.5	67.6	67.3	66.9	66.2	64.4	67.4	67.1	66.7	66.0	64.3	**66.7**

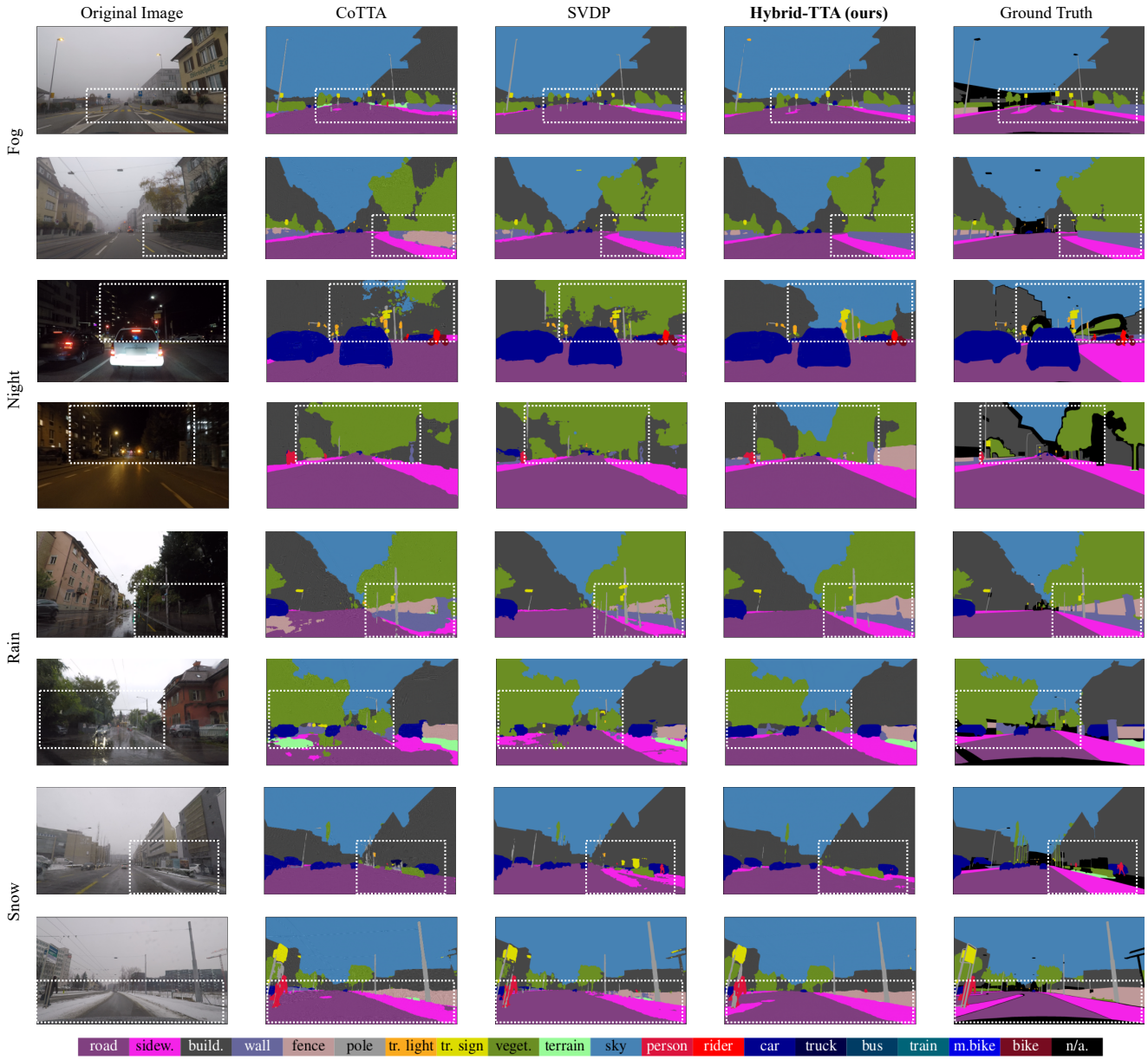


Figure 8. **Qualitative comparison of segmentation results on Fog, Night, Rain and Snow domains for Cityscapes-to-ACDC benchmark.**