Continual Adaptation of Vision Foundational Models for Semantic Segmentation in Adverse Weather - Supplementary

Nikhil Kumar Jangamreddy[1,2,3] Mahsa Baktashmotlagh[2] Chetan Arora[3]
[1]UQ-IIT Delhi Research Academy (UQIDRA), India [2]The University of Queensland, Australia
[3]Indian Institute of Technology Delhi, India

Comparison with Existing Test Time Adaptation (`TTA`) and Continual Test Time Domain Adaptation (`CTDA`) Methods. In Table 1, we compare `SegBridge` with existing `TTA` and `CTDA` methods. Note that existing `TTA` and `CTDA` methods are not compatible with `DINOv2` and `Mask2Former` since they typically rely on batch-normalization statistics and classification-style outputs. For fair comparison, we report `Mask2Former` +`DINOv2` (`M2FD2`) performance and then show the improvements achieved by applying our method on top of it. For existing `TTA` and `CTDA` methods, we report the results of existing approaches using source pre-trained models HGFormer [5] and DeepLabv3+ ResNet101 [2], both trained in the daytime. We use the DeepLabv3+ ResNet101 backbone in cases where approaches are specifically designed for CNN architectures. For example, TENT [20] is designed to adapt `BN` parameters, which are suitable only for CNN-based architectures. On the cityscapes-val dataset, the `mIoU` of the above-mentioned source pre-trained models is 80.4% and 78.5%, respectively. We report average `mIoU` across 10 iterations to demonstrate an overall improvement in semantic segmentation performance.

Supplementary Videos. All six MPEG-4 clips are located in `multimedia/` folder (720p / H.264).

Datasets Description.

Datasets. Our experiments are conducted on five benchmark datasets: C-Driving [10], IDD-AW [18], SHIFT [19], ACDC [17], and Cityscapes (Day) [4]. A brief overview of each dataset is provided below.

C-Driving [10]. The C-Driving benchmark is derived from the BDD100K dataset [23] and contains images under four adverse weather conditions: cloudy, rainy, snowy, and overcast. The validation split includes 346 cloudy, 215 rainy, 242 snowy, and 927 overcast scenes. For reporting, we use the C-Driving Rain validation subset.

IDD-AW [18]. The IDD-AW dataset consists of 5,000 finely annotated images representing complex driving environments in India. It captures diverse challenging scenarios such as rain, fog, snow, and low-light conditions. For evaluation, the IDD-AW validation set is employed.

SHIFT [19]. SHIFT provides a large-scale synthetic dataset of about 2.5M images covering multiple adverse scenarios.

ACDC [17]. The ACDC dataset contains 4,006 high-resolution images (1920×1080) collected under adverse weather conditions. It spans four domains: fog, night, snow, and rain. Each domain provides 400 labeled training images, around 100 validation images (106 in the case of night), and 500 unlabeled test samples.

Cityscapes (Day) [4]. The Cityscapes dataset offers 5,000 urban driving scenes captured in clear daytime conditions, with an image resolution of 2048×1024. The validation set includes 500 labeled samples, while the official test set contains 1,525 images for benchmarking.

Evaluation Criteria. Mean Intersection over Union (`mIoU`) is used as an evaluation criterion. Higher `mIoU` indicates better semantic segmentation predictions.

Related work: Multi-Prototype Learning. Infinite Mixture Prototypes [1] uses class-specific clusters to handle varying data complexities in few-shot learning by dynamically adjusting the number of clusters. `SegBridge` framework uses class-specific prompt embeddings within a continual learning framework to adapt to new domains while retaining previous knowledge using a greedy query replay buffer. IMP's embeddings are data-driven clusters, while SegBridge's are learnable prompts embeddings integrated into `SOTA` Vision Foundation Models. `SegBridge` offers a modular and parameter-efficient framework that enhances generalization. We also propose a greedy replay buffer to adapt to and retain knowledge across continuously shifting domains. By integrating Vision Foundation Models (e.g., `DINOv2`) with `Mask2Former`, `SegBridge` offers a modular and parameter-efficient framework that enhances generalization. The comparison results for (IDD-AW, C-Driving) datasets are: IMP: (58.4, 60.9) and `SegBridge` (ours): (69.8, 68.7).

Limitations. During our testing on unseen road scenes from YouTube videos, we observed that the segmentation of minority classes, such as traffic lights, is occasionally in-

Time	t ⟶																
Round	1				4				7				10				All
Model+Method	fog	night	rain	snow	fog	night	rain	snow	fog	night	rain	snow	fog	night	rain	snow	mean
HGFormer[5]	69.2	52.5	72.1	68.4	69.2	52.5	72.1	68.4	69.2	52.5	72.1	68.4	69.2	52.5	72.1	68.4	65.5
DeepLabv3+ ResNet101[2]	67.5	22.1	52.3	50.7	67.5	22.1	52.3	50.7	67.5	22.1	52.3	50.7	67.5	22.1	52.3	50.7	48.2
• TENT-continual [20]	69.2	23.4	53.5	51.3	65.2	17.4	50.9	49.9	61.5	15.5	48.9	48.4	59.9	14.4	48.2	48.1	45.4
• MEMO[25]	60.1	44.2	63.2	52.2	62.2	44.8	63.7	53.2	58.8	42.8	59.5	52.4	48.7	20.9	52.1	50.2	51.8
• EATA-continual[11]	69.5	28.9	56.2	55.1	67.9	28.1	55.1	53.4	65.2	26.3	51.9	52.1	65.1	25.9	53.1	52.1	50.4
• SAR[12]	68.8	24.2	53.4	51.1	69.2	25.5	54.9	52.8	70.5	27.2	55.1	53.4	71.1	28.9	56.5	56.2	51.2
• NOTE[7]	69.1	22.8	54.5	54.2	70.2	23.7	55.8	57.2	68.4	20.9	55.1	56.1	68.1	20.1	54.5	56.1	50.5
• RoTTA[24]	69.4	24.1	55.1	55.2	71.4	24.4	55.1	57.1	69.4	22.3	53.2	56.1	67.8	20.2	53.1	54.2	50.2
• DoSE[16]	73.4	26.4	57.7	57.9	76.5	27.9	60.4	60.3	76.1	28.2	61.1	60.4	76.8	30.9	63.8	60.7	56.2
• MECTA[8]	69.3	23.5	55.1	54.7	70.3	24.8	55.6	55.3	70.6	24.3	56.1	56.3	69.1	22.7	55.1	54.9	51.2
• RATP[9]	70.1	53.8	72.9	69.2	71.3	53.6	73.1	69.8	71.8	53.8	74.2	69.9	70.4	53.5	73.1	68.8	66.8
• RMT[6]	69.5	53.8	73.4	69.2	70.2	54.2	74.1	69.9	70.4	54.8	74.3	70.2	70.6	55.2	74.8	70.6	71.2
• MALL[15]	70.4	54.8	74.3	70.2	70.8	55.9	76.2	73.1	68.4	51.1	72.4	69.3	64.1	48.2	68.1	66.8	65.9
• CoTTA[21]	70.1	54.2	72.9	69.1	70.6	55.4	73.4	69.6	71.1	56.2	74.1	70.2	71.4	56.8	74.8	70.9	70.9
• ONDA[14]	71.1	54.8	58.4	51.8	73.2	57.8	56.8	55.1	68.9	55.7	53.8	56.9	66.2	51.4	53.9	54.2	68.3
M2FD2 [3, 13]	70.6	69.5	68.4	74.2	70.6	69.5	68.4	74.2	70.6	69.5	68.4	74.2	70.6	69.5	68.4	74.2	70.7
• **SegBridge (ours)**	**69.5**	**72.9**	**74.2**	**72.5**	**73.2**	**75.6**	**76.5**	**75.3**	**74.9**	**75.8**	**77.6**	**76.2**	**74.5**	**78.5**	**79.1**	**78.1**	**77.9**

Table 1. Comparison of SegBridge with existing TTA and CTDA methods on ACDC dataset, We use HGFormer[5] and DeepLabv3+ ResNet101[2] pre-trained on daytime as source pre-trained models. Results of our SegBridge are based on DINOv2 as pre-trained VFM. M2FD2 refers to Mask2Former with DINOv2 backbone.

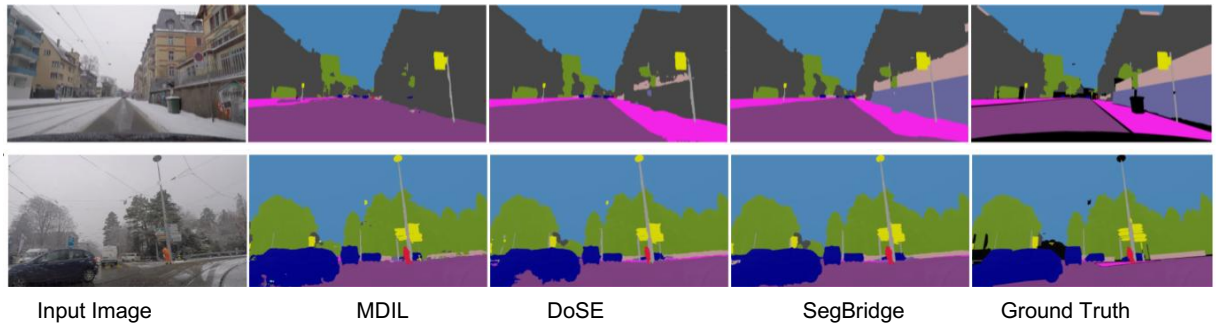

Figure 1. Qualitative visual comparison of proposed SegBridge approach with existing CLSS methods

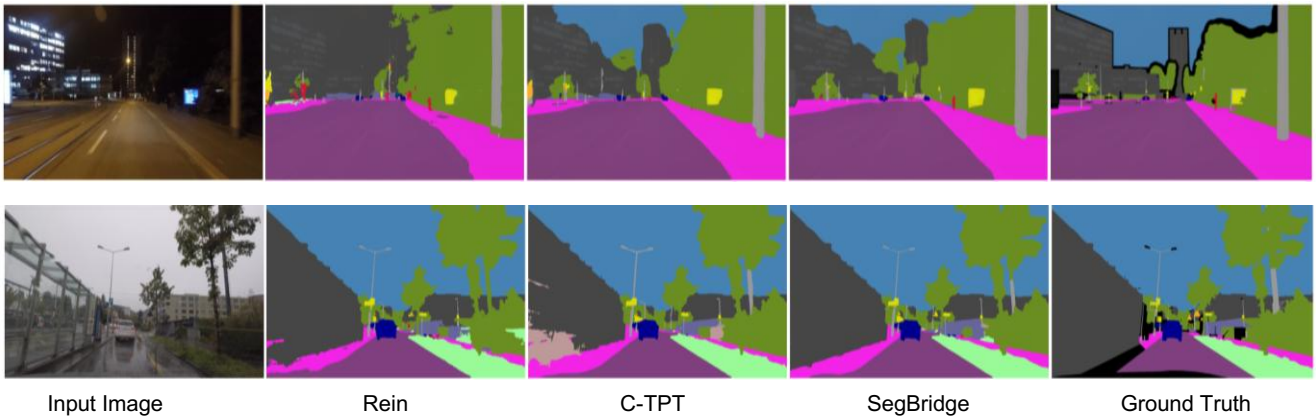

Figure 2. Qualitative visual comparison of proposed SegBridge approach with existing DG and PEFT methods

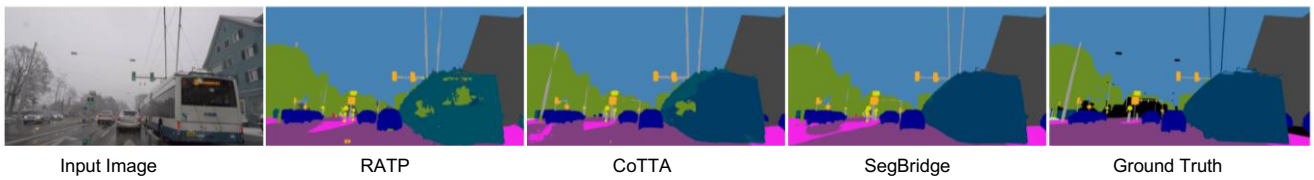

Figure 3. Qualitative visual comparison of proposed SegBridge approach with existing TTA and CTDA methods

ID	Description and filename
V1	Clear→Cloudy (ours vs. Rein) cloudy_comparison_with_rein.mp4
V2	Clear→Fog (robustness as visibility drops) fog_comparison_with_cotta.mp4
V3	Clear→Night (boundary discovery in low-light) night_comparison_with_c-tpt.mp4
V4	Clear→Overcast (VFM backbone under dull light) overcast_comparison_with_dinov2.mp4
V5	Clear→Rain (motion-blur mitigation) rain_comparison_with_mdil.mp4
V6	Clear→Snow (boundary recovery amid flakes) snow_comparison_with_soma.mp4

Table 2. Catalogue of supplementary videos.

accurate. However, due to their critical role in safety, ensuring accurate segmentation for minority classes is essential. Video outputs are provided in the supplementary.

Effect of Architectural Components. We perform a leave-one-out ablation to quantify the individual and synergistic benefits of the three modules that distinguish SegBridge from a straightforward VFM+Mask2Former baseline:

Per-Class IoU on ACDC. are reported in Table 4.

- **Visual Prompt Tuning (VPT).** Injects lightweight prompt tokens into the frozen decoder, enabling fast test-time plasticity without full fine-tuning.
- **Boundary-Guided Query Discovery (BGQD).** Refines object queries using high-confidence edge cues, yielding crisper masks under blurred or low-contrast conditions.
- **Mixture-of-Experts (MoE) Gate.** Dynamically routes queries to domain-specialised prompt experts, mitigating catastrophic forgetting as domains evolve.

Table 5 reports mIoU on the two most challenging datasets (C-Driving and IDD-AW) after ten CTDA iterations. Starting from a frozen Mask2Former, VPT alone already gives a +2.1 / +1.9 mIoU gain, confirming that prompt tokens are a parameter-efficient alternative to decoder retraining. Adding BGQD delivers a further +2.1 / +1.8 mIoU by sharpening object boundaries, while the MoE gate contributes an additional +1.2 / +1.0. Stacking all three components yields a 6–7 % mIoU improvement over the baseline and establishes new state-of-the-art performance on both datasets.

The full model adds fewer than **2.3 M** trainable parameters (0.8 % of DINOv2) and increases test-time latency by only **0.13 s / image**, making it suitable for real-time deployment in adverse driving scenarios.

Incorporation of VFMs. To extract image features as input for the pixel decoder and transformer decoder, the Mask2Former pixel decoder, transformer decoder receives feature maps from Vision Foundation Models (VFMs) such as DINOv2 and EVA02. These feature maps are derived from multiple predefined layers, including the 7th, 11th, 15th, and 23rd layers, similar to Rein [22], providing rich spatial and semantic information essential for downstream processing.

Hyperparameters used:. We initialise the domain-specific prompt embeddings for each domain with a dimension of $D = 256$ and pass them through a Visual Prompt Tuning (VPT) adapter comprising $L = 9$ transformer layers. The Mixture-of-Experts gating network is implemented as a lightweight MLP with two hidden layers, each containing 256 neurons. Following Mask2Former, we use $N = 100$ base object queries and $N_d = 50$ domain-specific prompt tokens, resulting in 8 learnable prompt embeddings per class. To mitigate forgetting, a greedy query replay buffer of size $n = 10$ is maintained. During continual adaptation we optimise only the prompts, VPT projection, and MoE gate with the Adam optimiser (lr $= 1\times10^{-3}$, momentum $= 0.9$). The loss trade-off is governed by $\lambda_{\text{NC}} = 10^{-1}$, while the temperature and confidence thresholds are fixed to $\tau = 0.07$ and $\theta = 0.5$, respectively. Edge-aware anchors are obtained with a Canny detector using low/high thresholds $(30, 150)$, and all Vision Foundation Model (VFM) weights remain frozen during inference.

Qualitative Visual Comparison. We report the qualitative visual results of SegBridge, comparing its performance with current state-of-the-art methods. Results are reported in Figure 1, Figure 2 and Figure 3 respectively.

Video Outputs. We provide video outputs in the supplementary material to compare the proposed SegBridge approach with the SOTA CTDA, DG, and CS methods.

Ablation Study on LoRA Rank. We perform an ablation study on the ACDC dataset to analyze the impact of the LoRA rank r on the SegBridge framework. Keep-

Time	t ⟶																
Round	1				4				7				10				All
Method	rain	cloud	overcast	snow	rain	cloud	overcast	snow	rain	cloud	overcast	snow	rain	cloud	overcast	snow	mean
M2FD2 [3, 13]	62.9	60.9	60.4	61.2	62.9	60.9	60.4	61.2	62.9	60.9	60.4	61.2	62.9	60.9	60.4	61.2	61.3
• **SegBridge (ours)**	**60.1**	**65.3**	**69.7**	**71.2**	**61.9**	**66.3**	**71.2**	**72.3**	**62.2**	**68.1**	**72.3**	**73.5**	**64.8**	**68.9**	**73.5**	**74.7**	**68.8**

Table 3. Results of SegBridge framework using pre-trained DINOv2 [13] on C-Driving dataset. The results show the performance for different weather conditions (rain, cloud, overcast, and snow) over 10 iterations. M2FD2 refers to Mask2Former with DINOv2 backbone.

Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
SegBridge	93.2	74.5	83.1	43.8	34.2	66.7	59.4	60.3	87.9	56.0	95.4	73.7	58.9	92.3	72.5	79.6	62.4	46.2	65.3
CoTTA	90.1	68.2	79.3	41.2	30.9	60.4	51.3	55.8	85.1	49.7	92.2	67.4	51.6	90.1	69.4	75.7	59.8	41.7	58.6

Table 4. IoU (%) for each of the 19 Cityscapes classes after 10 CTDA rounds.

Configuration	**C-Driving**	**IDD-AW**
Frozen VFM +Mask2Former	61.3	59.2
+ VPT	65.2	61.3
+ BGQD	67.3	63.1
+ MoE	66.4	63.2
All components (ours)	**68.7**	**65.1**

Table 5. Ablation of architectural components (mIoU, %).

DINOv2 Variant	**Average mIoU (%)**
DINOv2/ViT-S/14	65.8
DINOv2/ViT-B/14	69.5
DINOv2/ViT-L/14 (DINOv2-L)	74.2

Table 6. Ablation study of different DINOv2 variants on SegBridge performance. We report the Average mIoU (%) on the ACDC dataset. Larger DINOv2 variants, such as ViT-L/14, demonstrate improved performance.

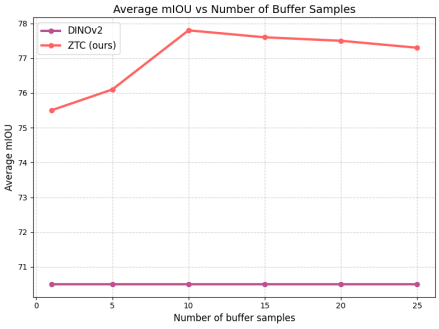

Figure 4. Impact of buffer sample size on average mIOU.

ing all other hyperparameters fixed, we vary r and report the average mIoU across all four adverse-weather domains. As shown in Tab. 7, rank ($r = 8$) provides the best performance, while both smaller and larger ranks lead to under- and over-parameterization effects, respectively.

LoRA Rank r	**Average mIoU**
2	62.1
4	64.3
8	**68.4**
16	66.9
32	64.1.x

Table 7. Ablation study of LoRA rank on the ACDC dataset. A moderate rank ($r = 8$) yields the highest average mIoU.

Ablation on different variants of VFMs. To assess the impact of different DINOv2 variants on proposed SegBridge performance, we perform an ablation study using various configurations of the DINOv2 model as the base pre-trained Visual Foundation Model (VFM). The results, presented in Table Figure 6, demonstrate the Average mIoU (%) achieved on the ACDC dataset for each variant.

Ablation Study: Impact of Sample Size in Greedy Query Replay Buffer. To analyze the impact of the sample size in the Greedy Query Replay Buffer on model performance, we conduct an ablation study by varying the number of samples in the buffer. We use DINOv2 as our base pre-trained VFM model and report results on ACDC dataset. A sample corresponding to the image embedding and domain-specific prompt embedding. The results are reported in Figure 4.

Affect of the Order Sequence. To analyze the impact of the order sequence of domains on the segmentation performance of the SegBridge framework, we alter the domain order and report the results in Table 3.

Results on SHIFT dataset. are reported in Tab. 8.

Time	$t \longrightarrow$																
Round	1				4				7				10				All
Model+Method	fog	night	rain	overcast	fog	night	rain	overcast	fog	night	rain	overcast	fog	night	rain	overcast	mean
DINOv2 [13]	67.9	62.3	63.8	62.6	67.9	62.3	63.8	62.6	67.9	62.3	63.8	62.6	67.9	62.3	63.8	62.6	64.2
• **SegBridge (ours)**	**68.9**	**64.1**	**65.1**	**66.4**	**70.1**	**67.3**	**66.2**	**68.3**	**71.1**	**68.4**	**68.2**	**69.7**	**72.4**	**70.1**	**69.9**	**70.2**	**68.4**

Table 8. Results of SegBridge framework on SHIFT dataset. Results of SegBridge are based on DINOv2 as pre-trained VFM.

References

[1] Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. Infinite mixture prototypes for few-shot learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 232–241. PMLR, 2019. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 4

[4] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[5] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15413–15423, 2023. 1, 2

[6] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 2

[7] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[8] Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. Mecta: Memory-economic continual test-time model adaptation. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[9] Chenxi Liu, Lixu Wang, Lingjuan Lyu, Chen Sun, Xiao Wang, and Qi Zhu. Deja vu: Continual model generalization for unseen domains. *arXiv preprint arXiv:2301.10418*, 2023. 2

[10] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[11] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The International Conference on Machine Learning*, 2022. 2

[12] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023. 2

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 5

[14] Theodoros Panagiotakopoulos, Pier Luigi Dovesi, Linus Härenstam-Nielsen, and Matteo Poggi. Online domain adaptation for semantic segmentation in ever-changing conditions. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[15] Nikhil Reddy, Abhinav Singhal, Abhishek Kumar, Mahsa Baktashmotlagh, and Chetan Arora. Master of all: Simultaneous generalization of urban-scene segmentation to all adverse weather conditions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 51–69. Springer, 2022. 2

[16] Nikhil Reddy, Mahsa Baktashmotlagh, and Chetan Arora. Domain-aware knowledge distillation for continual model generalization. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 685–696, 2024. 2

[17] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. *arXiv preprint arXiv:2104.13395*, 2021. 1

[18] Furqan Ahmed Shaik, Abhishek Reddy, Nikhil Reddy Billa, Kunal Chaudhary, Sunny Manchanda, and Girish Varma. Idd-aw: A benchmark for safe and robust segmentation of drive scenes in unstructured traffic and adverse weather. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4614–4623, 2024. 1

[19] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 1

[20] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2

[21] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai.

Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7191–7201, 2022. 2

[22] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28619–28630, 2024. 3

[23] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[24] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. 2

[25] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022. 2