A. Supplementary Material

In Section A.1, we provide a description of our new dataset. The transition effects of agents moving between cells are briefly studied in Section A.2. We also present additional experiments on cyclic domain shifts in Section A.3.

A.1. DADE Dataset

To study our new Multi-Stream Cellular Test-Time Adaptation (MSC-TTA) setup and evaluate the performance of our real-time method, we need a dataset that meets the four following criteria. **(1) Multi-agent long videos:** the dataset should consist of long video sequences captured by multiple agents operating within the same dynamic environment. **(2) Environment division:** the environment should be heterogeneous or dynamic to be spatially and/or temporally divided into cells, *e.g.*, encompassing a variety of driving locations, such as rural, urban, and highway settings, or a broad spectrum of weather conditions, including, *e.g.*, day, night, clear, rainy, and foggy scenarios. **(3) Cell connection:** each agent's connection to a cell should be precisely estimated, for instance using GNSS (Global Navigation Satellite System) coordinates for the location, or a weather service for the weather conditions. **(4) Available ground truths:** for evaluation purposes, we need to have access to ground-truth annotations for our semantic segmentation task. Unfortunately, publicly-available datasets do not meet these criteria. Existing datasets, such as [7, 32, 41], typically feature short video sequences, lack multi-agents, or often do not include ground-truth annotations or a diverse range of weather conditions. While the SHIFT dataset [41] contains varying weather conditions and ground truths, it is not a multi-agent dataset and its average sequence length is under 160 s, which is too short for evaluating the long term impact of our method.

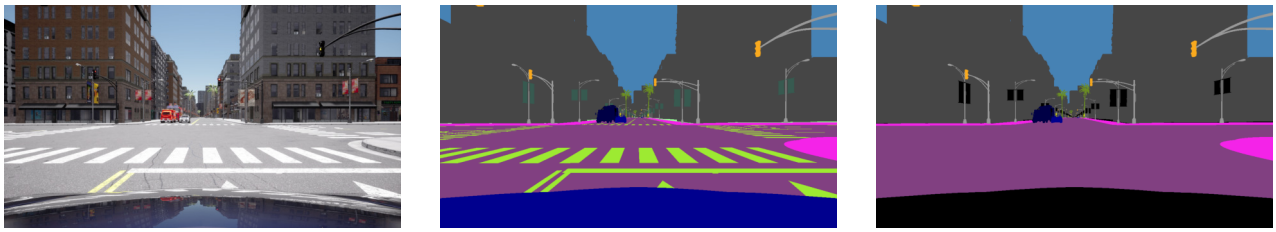
Therefore, we generated and will publicly release our own Driving Agents in Dynamic Environments (DADE) dataset, meeting all the above criteria. To have access to ground-truth annotations and precisely control the environment, we choose the CARLA simulator [9] (version 0.9.14) to generate the dataset. We synchronize and calibrate all sensors and register the semantic segmentation ground truths. Our dataset is acquired using the recent `Town12` map that offers several visually distinct locations and fine-grained control over the weather. Our simulation showcases several agents, in our case, ego vehicles, on which a camera is attached at the front, filming its front view (in a "Cityscapes" fashion), as shown in Figure 6. We collect the video sequences taken by an RGB camera, the semantic segmentation ground-truth masks, the GNSS position of each agent in the simulation as well as the overall weather information. All signals are acquired at the framerate of 1 frame per second, with a high-resolution (HD) definition.

Table 3. **Comparison of class definition** between Cityscapes [7], CARLA [9], and our DADE dataset. Our DADE dataset takes the intersection of the class definition between Cityscapes and CARLA. The classes not included in the intersection are projected to the "unlabeled" class, except for "road line" which is projected to "road". The classes used in training and evaluation for DADE are the same as the ones of Cityscapes.

Cityscapes			CARLA	DADE		
name	**training**	**evaluation**	**name**	**name**	**training**	**evaluation**
unlabeled			unlabeled	unlabeled		
static			static	static		
dynamic			dynamic	dynamic		
ground			ground	ground		
road	✓	✓	road	road	✓	✓
sidewalk	✓	✓	sidewalk	sidewalk	✓	✓
rail track			rail track	rail track		
building	✓	✓	building	building	✓	✓
wall	✓	✓	wall	wall	✓	✓
fence	✓	✓	fence	fence	✓	✓
guard rail			guard rail	guard rail		
bridge			bridge	bridge		
pole	✓	✓	pole	pole	✓	✓
traffic light	✓	✓	traffic light	traffic light	✓	✓
traffic sign	✓	✓	traffic sign	traffic sign	✓	✓
vegetation	✓	✓	vegetation	vegetation	✓	✓
terrain	✓	✓	terrain	terrain	✓	✓
sky	✓	✓	sky	sky	✓	✓
person	✓	✓	person	person	✓	✓
rider	✓	✓	rider	rider	✓	✓
car	✓	✓	car	car	✓	✓
truck	✓	✓	truck	truck	✓	✓
bus	✓	✓	bus	bus	✓	✓
motorcycle	✓	✓	motorcycle	motorcycle	✓	✓
bicycle	✓	✓	bicycle	bicycle	✓	✓
ego vehicle			ego vehicle	ego vehicle		
rectification border			other			
out of roi			road line			
parking			water			
tunnel						
caravan						
trailer						
train						

To align our dataset with current benchmarks in the semantic segmentation field, we generated two versions of the semantic segmentation ground truths in the dataset: (1) the ones directly collected from the CARLA simulator (including 29 semantic classes), and (2) the intersection of the semantic classes available in CARLA and the semantic classes from the Cityscapes [7] dataset (including 33 different semantic classes). For consistency with previous works, we choose the later version in our experiments, as most state-of-the-art models for semantic segmentation in driving environments are trained on Cityscapes. Nevertheless, the discrepancies between the two versions are minimal and could be interchanged depending on the user's preferences. Table 3 provides the complete comparison between the semantic segmentation classes of the Cityscapes dataset, the CARLA simulator and our DADE dataset. Figure 6a shows an RGB image alongside the two versions of the semantic segmentation ground-truth masks (Figure 6b and Figure 6c). As can be seen, the "road line" class in the CARLA labels, visible in Figure 6b, does not exist in the Cityscapes labels. Also, the "car hood" is ignored (indicated by black pixels) in the second version.

To study different cell divisions of the environment, our DADE dataset is composed of two parts. The first part,



(a) (b) (c)

Figure 6. **Comparison between the real ground truth and the ground truth used in our experiments.** (a) An RGB image with (b) its corresponding semantic segmentation ground truth from CARLA and (c) the semantic segmentation ground truth that we used to evaluate our method. The black pixels in image (c) correspond to ignored classes or regions, such as the hood of the ego vehicle.

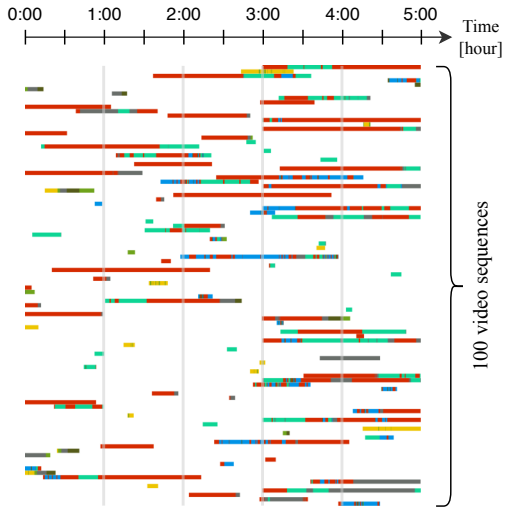


Figure 7. **Location of each agent** for the 100 sequences of the DADE-static dataset. The color of the line corresponds to the location of the agent for each sequence at a given time: forest, countryside, rural farmland, highway, low density residential, community buildings, and high density residential. We can see that the sequences are evenly distributed across the entire 5-hour time frame.

DADE-static, is acquired with static weather conditions (clear day) and contains 100 video sequences, as shown in Figure 7. The second part, *DADE-dynamic*, is acquired with varying weather conditions (ranging from day to night, with clear, rainy or foggy conditions) and contains 300 video sequences, as shown in Figure 8. For both parts, each sequence is acquired by one agent (one ego vehicle) running for some time within a 5-hour time frame, amounting to a total of 990k frames for the entire dataset. In Figure 9, we show a top view of the various locations in the `Town12` map of the CARLA simulator in which the agents evolve, namely forest, countryside, rural farmland, highway, low density residential, community buildings, and high density residential. Images captured in each location can be seen in Figure 10. Finally, Figure 11 illustrates the 6 different weather conditions, in the high density residential location, encountered in the DADE-dynamic dataset.

Let us note that due to the limitations of the CARLA simulator running the `Town12` map, there are no pedestrians on the streets, only vehicles such as cars, motorcycles, bicycles or trucks. Also, the quantity of vehicles (traffic) is independent on the location. The vehicles spawned in the map move randomly through the seven locations. Finally, the different sequences are collected sequentially. In the following, we provide some statistics about both parts of our DADE dataset.

A.1.1 DADE-static dataset

This first part of our dataset is composed of 100 sequences, acquired in the `Town12` map of CARLA with a static clear sunny weather during the day. Each sequence contains between 271 and 7,200 frames acquired at 1 fps, for total of 270,527 frames, amounting to more than 75 hours of video. In Figure 12a, we show the distribution of the sequence length for the 100 sequences. As can be seen, our dataset contains a lot of short and long sequences, with an average sequence length of 45 minutes. We also show, in Figure 7, the locations of the 100 agents over time. The colors correspond to the locations in which the agents are evolving (see Figure 9). We can see that, for most sequences, the agents evolve through several locations, and that the start and end times vary significantly from one agent to another.

Figure 13a provides a more detailed analysis of each agent's location over time. Particularly, it shows that there is a high imbalance between the locations, which is expected in real-world scenarios. For instance, it is realistic to encounter much more vehicles in city centers than in the countryside. Table 4 summarizes those values and splits the number of images acquired during the two first hours (used for pretraining) and the three last hours (used for testing). Interestingly, data originating from the high density residential location constitute over half of our DADE-static dataset. We can also see that, during the first two hours, over a thousand images are collected in each location, con-

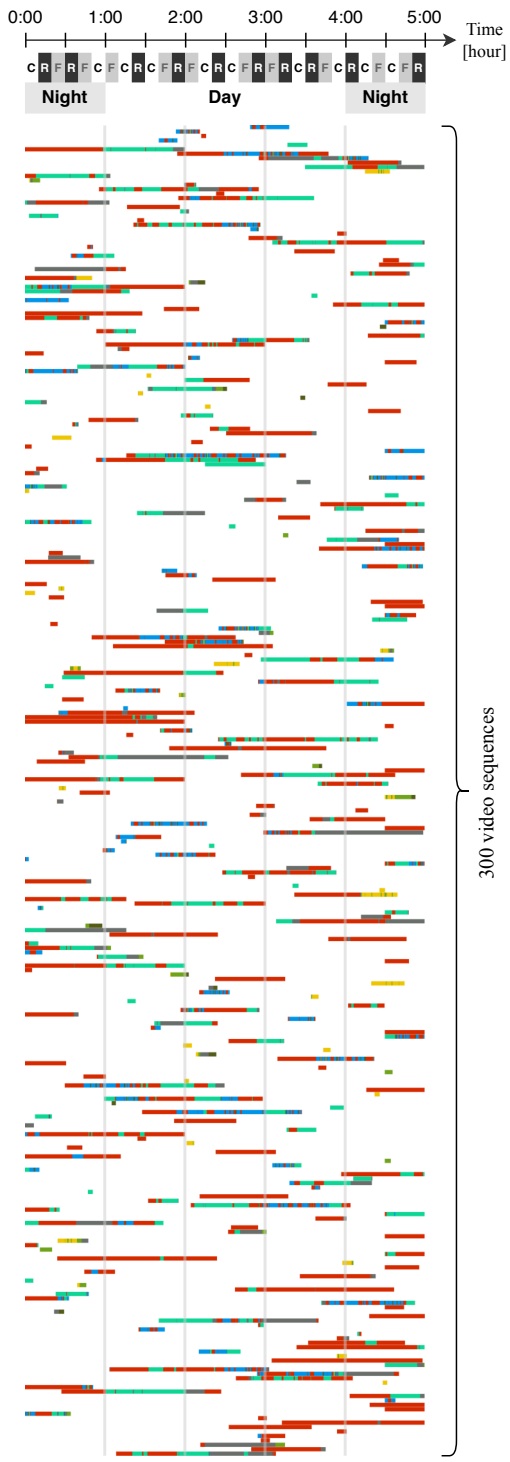


Figure 8. **Location, weather, and daylight conditions of each agent** for the 300 sequences of the DADE-dynamic dataset. C, R and F respectively correspond to clear, rainy and foggy weathers, and night/day represent the daylight conditions. The color of the line corresponds to the location of the agent for each sequence at a given time: forest, countryside, rural farmland, highway, low density residential, community buildings, and high density residential.

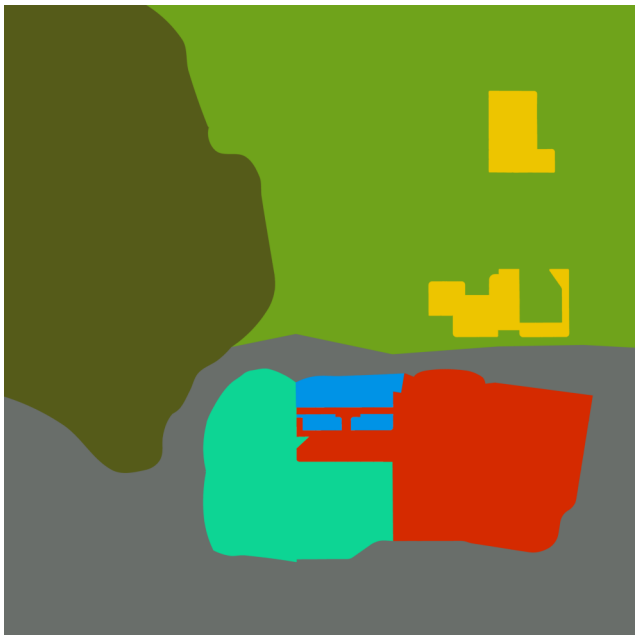


Figure 9. **Locations of the `Town12` map in CARLA.** The figure provides the location in which the vehicle is depending on its x and y coordinates. The different locations are the following: forest, countryside, rural farmland, highway, low density residential, community buildings, and high density residential.

stituting a sufficient pretraining set.

Table 4. **Number of images per location** within the DADE-static dataset during the pretraining time (two first hours), the test time (three last hours), and the overall time (the five hours), as well as the proportion of images originating from each location in comparison to the entire dataset.

Location	Pretraining (2 hours)	Testing (3 hours)	Overall (5 hours)	Proportion of the entire dataset
Forest	2,176	2,796	4,972	1.84%
Countryside	2,442	1,215	3,657	1.35%
Rural farmland	3,608	6,089	9,697	3.58%
Highway	7,018	19,159	26,177	9.68%
Low density residential	11,187	36,658	47,845	17.69%
Community buildings	2,357	20,404	22,761	8.41%
High density residential	50,034	105,384	155,418	57.45%
Total	**78,822**	**191,705**	**270,527**	**100%**

A.1.2 DADE-dynamic dataset

This second part of our dataset is acquired during a period of time of 5 hours with varying weather conditions as shown in Figure 8. Particularly, it is composed of 300 sequences containing between 188 and 7,200 frames acquired at 1 fps, for a total of 719,742 frames or 200 hours of videos. Figure 12b shows the distribution of the sequence length. It can be noted that the distribution follows the same trend as for DADE-static, with a similar average sequence length of

Figure 10. **Examples of the different locations in our dataset.** We define 7 different locations based on the GNSS data and show some images captured by the agent in each location. From left to right, we display the name of the location, an overview of the location, and six images captured by agents.

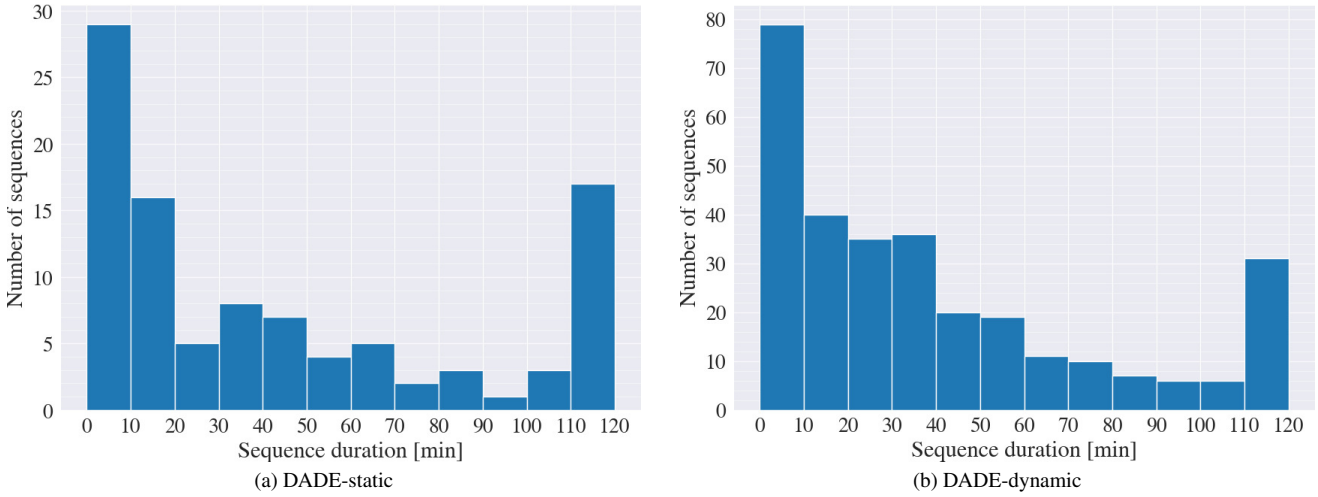
Figure 11. **Examples of the different weather and daylight conditions in our dataset.** The images show the 6 different weather and daylight conditions in the same high density residential location. The images correspond respectively, from left to right and top to bottom, to clear day, clear night, rainy day, rainy night, foggy day, and foggy night.

40 minutes. To generate various weather conditions, we dynamically change the weather parameters over time in the same way for the entire map, *i.e.*, that the weather condition is the same for all agents at a given time. We change the weather condition every 10 minutes arbitrarily between clear, rainy, and foggy weathers, with a smooth transition on the weather parameters during 10 seconds. For the daylight conditions, we choose to start the simulation during night time and let the sun rise after one hour, finally setting in the last hour. The 5 hours are thus composed of a total of approximately 2 hours of night conditions and 3 hours of day conditions. During the entire simulation, there are 4 periods of 10 minutes for each weather condition (*i.e.* clear, rainy, and foggy) during the night and 6 during the day as shown in Figure 14a. To visualize the transitions, Figure 14b zooms in on the first thirty minutes where the simulation changes from clear, then rainy, and finally foggy weathers.

Table 5 provides a summary of the number of images for each location, weather, and daylight conditions during the two first hours (used for pretraining) and the three last hours (used for testing). We can see that the proportion of images in each location is similar to the one of our DADE-static dataset. However, upon further division based on both weather and daylight conditions, we can see a significant decrease in the number of images for each cell. Notably, this division results in the absence of pretraining data for the clear day weather condition in the countryside location. Finally, Figure 13b shows the number of agents in each location. The color of the plots corresponds to the color code of the location in the `Town12` map (see Figure 9). As can be seen, there is also a high imbalance between the different locations.

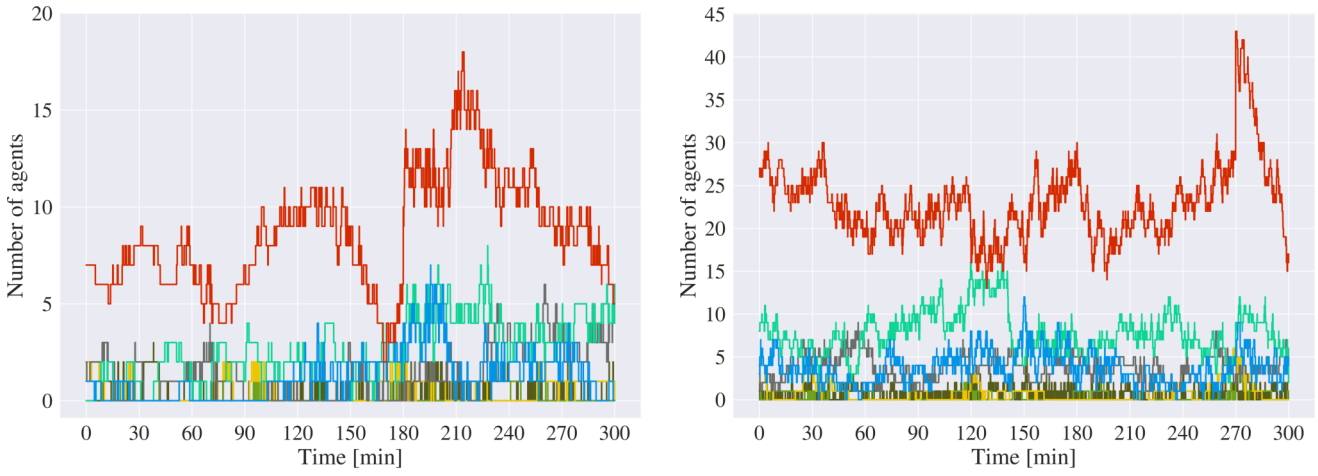
Table 5. **Number of images per location** within the DADE-dynamic dataset during the pretraining time (two first hours), the test time (three last hours), and the overall time (the five hours), as well as the proportion of images originating from each location in comparison to the entire dataset.

Location	Pretraining (2 hours)	Testing (3 hours)	Overall (5 hours)	Proportion of the entire dataset
Forest	3,174	5,973	9,147	1.27%
Clear night	845	695	1,540	
Rainy night	303	477	780	
Foggy night	585	602	1,187	
Clear day	381	1,467	1,848	
Rainy day	572	1,675	2,247	
Foggy day	488	1,057	1,545	
Countryside	3,525	4,283	7,808	1.09%
Clear night	279	1,247	1,526	
Rainy night	1,137	130	1,267	
Foggy night	887	795	1,682	
Clear day	0	194	194	
Rainy day	1,020	1,312	2,332	
Foggy day	202	605	807	
Rural farmland	4,605	9,242	13,847	1.92%
Clear night	736	2,631	3,367	
Rainy night	1,134	265	1,399	
Foggy night	2,059	2,699	4,758	
Clear day	221	1,268	1,489	
Rainy day	418	926	1,344	
Foggy day	37	1,453	1,490	
Highway	27,573	40,275	67,848	9.43%
Clear night	4,676	4,878	9,554	
Rainy night	4,809	4,508	9,317	
Foggy night	6,052	4,876	10,928	
Clear day	3,533	7,575	11,108	
Rainy day	4,235	9,757	13,992	
Foggy day	4,268	8,681	12,949	
Low density residential	56,108	84,990	141,098	19.60%
Clear night	7,348	9,214	16,562	
Rainy night	6,957	7,381	14,338	
Foggy night	7,673	8,190	15,863	
Clear day	11,486	22,607	34,093	
Rainy day	11,736	17,570	29,306	
Foggy day	10,908	20,028	30,936	
Community buildings	23,965	42,205	66,170	9.19%
Clear night	3,648	4,984	8,632	
Rainy night	3,746	3,532	7,278	
Foggy night	3,386	4,121	7,507	
Clear day	4,838	9,096	13,934	
Rainy day	4,210	9,539	13,749	
Foggy day	4,137	10,933	15,070	
High density residential	164,708	249,116	413,824	57.50%
Clear night	26,627	38,134	64,761	
Rainy night	31,006	25,618	56,624	
Foggy night	28,064	33,440	61,504	
Clear day	26,260	48,795	75,055	
Rainy day	26,830	53,676	80,506	
Foggy day	25,921	49,453	75,374	
Total	**283,658**	**436,084**	**719,742**	**100%**



(a) DADE-static (b) DADE-dynamic

Figure 12. **Distribution of sequence lengths** for (a) the DADE-static dataset and (b) the DADE-dynamic dataset. The DADE-static and DADE-dynamic datasets have respectively an average sequence length of 45 and 40 minutes, with durations ranging from a few minutes to two hours.



(a) DADE-static (b) DADE-dynamic

Figure 13. **Number of agents per location** over time for (a) the DADE-static dataset and (b) the DADE-dynamic dataset. The colors of the plots correspond to the location: forest, countryside, rural farmland, highway, low density residential, community buildings, and high density residential. The same trends can be observed in both datasets, with three times as many agents in DADE-dynamic as in DADE-static. Note that there is at all time at least one agent in the high density residential location in both dataset and in the low density residential in the DADE-dynamic. Conversely, forest, countryside, and rural farmland locations exhibit the least agent presence, often remaining empty of agents for extended periods.

A.2. Analysis of transiting agents

In this section, we provide insights about the transition of agents between cells. Particularly, we study the evolution of the performance of the models around cell transitions. Figure 15 shows the mean performance of the agents transiting from one cell to another, *e.g.*, from a specific location to another, or from a weather condition to another. As can be seen, after the transition, the baseline method experiences a decrease in performance, which remains low for a long period of time. Contrarily, our method is able to recover much faster thanks to the switch between the cell-specific models. It is also interesting to observe that for both the MSC-OL and MSC-TTA setups, a temporary drop of mIoU score occurs right before a transition. This is probably due to the fact that a vehicle approaching another cell may already see content from an adjacent cell while performing the task with the previous cell's model. For instance, a vehicle approaching the city center may record an image with its frontal camera showing the city center while still being

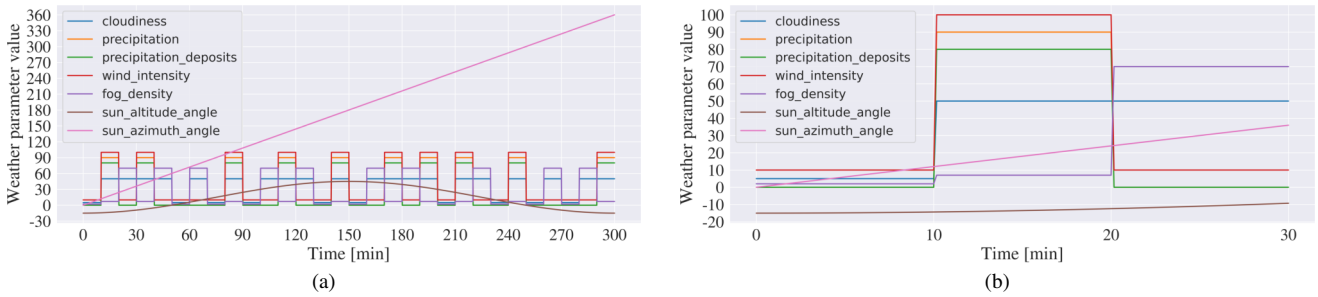


Figure 14. **Evolution of the weather parameters over time.** The weather switches arbitrarily every 10 minutes between clear, rainy, and foggy weathers with a smooth transition of 10 seconds. The parameters for the daylight conditions are related to the sun position, *i.e.* sun altitude angle and sun azimuth angle, which vary smoothly over time, respectively between -15 and 45 degrees, and between 0 and 360 degrees. We consider that it is night time during the first hour and the last hour, when the sun altitude is below 5 degrees, and that it is the day in between during three hours, *i.e.*, when the sun's altitude is over 5 degrees. In total, the clear, rainy, and foggy weathers each occur 10 times; 4 times during the night and 6 times during the day, as shown in (a). (b) zooms in on the first thirty minutes. During the first ten minutes, the weather is clear, then there is a smooth transition of 10 seconds towards a rainy weather and, finally, after 20 minutes there is again a smooth transition of 10 seconds towards a foggy weather.

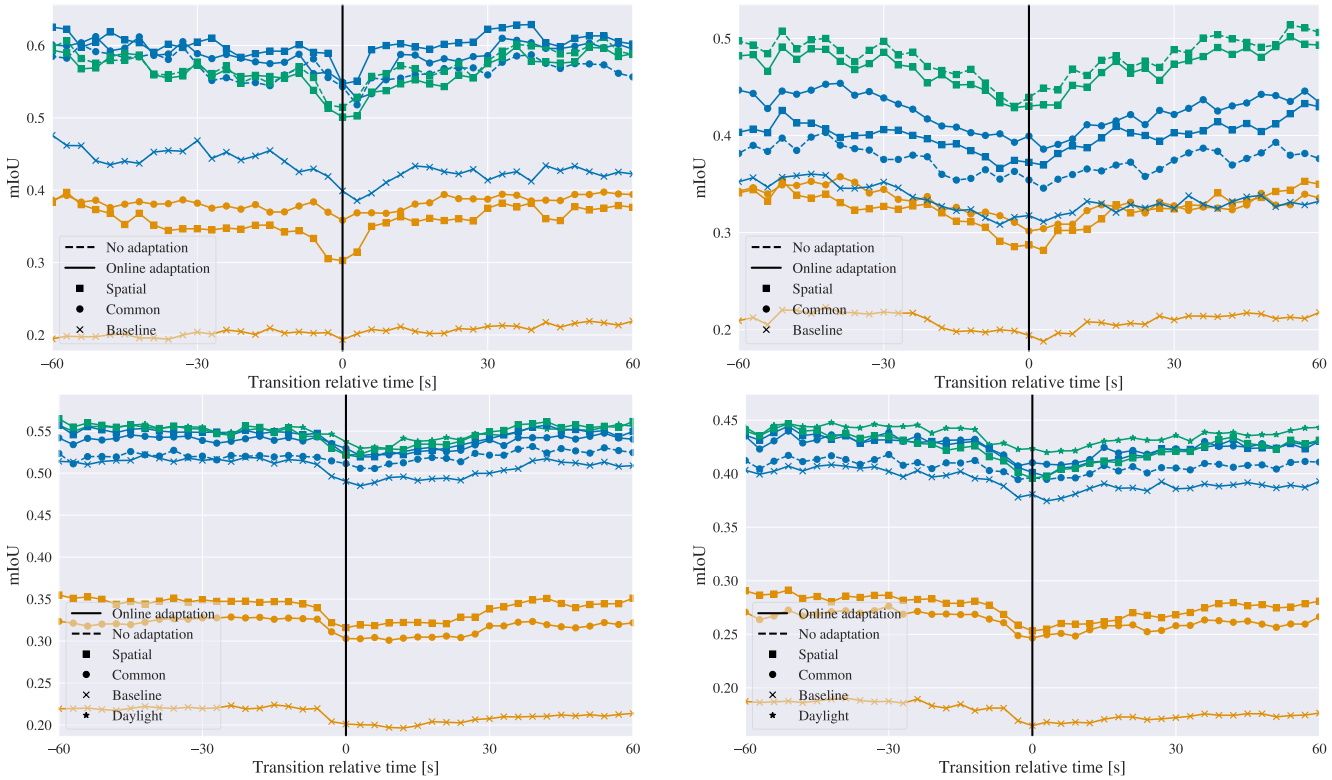


Figure 15. **Fleet performance around cell transitions on DADE-static dataset (top) and DADE-dynamic dataset (bottom).** Comparison of the performance in the MSC-OL setup (left) and the MSC-TTA setup (right) of our method (best settings) with the baseline for each pretraining (*Scratch*, *General*, and *Cell*). Confusion matrices of each frame are aggregated using a sliding windows of 3 seconds. The results are shown 30 seconds before and after any cell transition that the agents encounter during the 3 hours of testing.

registered in another location (*e.g.*, the countryside or the highway). This means that the agent will use the wrong cell model to analyze the environment. In future work, we aim to address this issue by proposing a model that automatically recognizes the cell, rather than relying on predefined rules.

Table 6. Comparison of our MSC-TTA method with a frozen teacher, a frozen student, the *Baseline* [6], and *Baseline*+MIR [19], on our DADE datasets and the dataset of Houyon *et al*. [19].

mIoU-I	DADE-S	DADE-D	Houyon [19]
Teacher ❄	.668	.611	/
Student ❄	.214	.159	/
Baseline [6]	.274	.212	.234
Baseline+MIR [19]	.181	.147	.256
Ours	**.362**	**.312**	**.277**

A.3. Experiments on cyclic domain shifts

In Table 6, we present additional experiments on the dataset and the best method proposed by Houyon *et al*. [19], namely *Baseline*+MIR, alongside the performance of the frozen teacher and student trained on the same set (namely, Cityscapes). The dataset and method [19] are specifically tailored for cyclic domain shifts. The two first columns are reported from Tables 1 and 2, for the 3 hours test sets.

Notably, our method demonstrates superior performance on DADE-static, DADE-dynamic, and the cyclic dataset of Houyon *et al*. [19]. Furthermore, we see that the *Baseline*+MIR performs worse than the baseline for DADE-static and DADE-dynamic, while it performs better on the cyclic dataset of Houyon *et al*. [19].

The results also demonstrate that our method exhibits an expected performance deficit relative to the teacher, while consistently outperforming the student. The teacher is a state-of-the-art semantic segmentation model (namely, SegFormer trained on Cityscapes [7]) and thus exhibits great performance on our DADE datasets. However, the emphasis on achieving the best possible performance often comes with increased complexity and overlooks the critical real-time aspect. The frame rate of SegFormer is approximately 2 frame per second, it is thus far from being real time. Our proposed method aims at mimicking the performance of available teacher models while reducing computational power and battery usage, thereby bringing state-of-the-art performance at a higher frame rate.

The frozen student is trained on the Cityscapes [7] dataset with an initial learning rate of 10^{-4} using the Adam optimizer for 45 epochs, reducing the learning rate by a factor of 10 every 15 epochs, a cross-entropy loss function, and a batch size of 8. To match the dimension of the images in the DADE datasets, images from Cityscapes were resized to 720x1440 to keep the same ratio, then cropped to 720x1280.