Supplementary Material

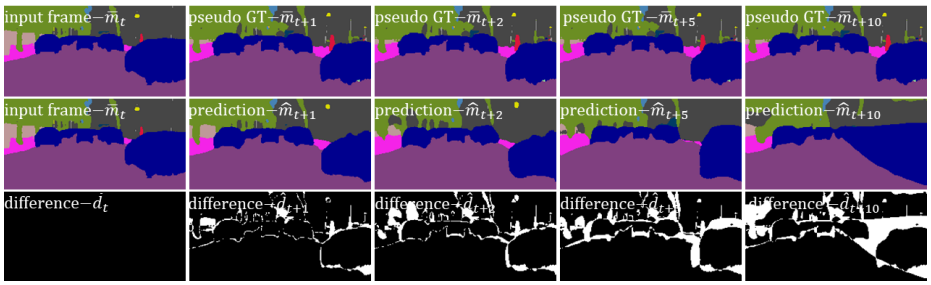

Fig. 1: Output predictions for a sequence of the BDD100K validation split, $\mathcal{D}_{\text{val}}^{\text{BDD-MOTS}}$. The top row depicts the pseudo ground truth $\overline{\mathbf{m}}_t$, $\overline{\mathbf{m}}_{t+1}$, $\overline{\mathbf{m}}_{t+2}$, $\overline{\mathbf{m}}_{t+5}$, $\overline{\mathbf{m}}_{t+10}$ generated by `PSANet`. In the middle row, we show the input semantic segmentation $\overline{\mathbf{m}}_t$ along with the predictions $\hat{\mathbf{m}}_{t+1}$, $\hat{\mathbf{m}}_{t+2}$, $\hat{\mathbf{m}}_{t+5}$, $\hat{\mathbf{m}}_{t+10}$ from the prediction network. The bottom row portrays the absolute difference $\hat{\mathbf{d}}_{t+1}$, $\hat{\mathbf{d}}_{t+2}$, $\hat{\mathbf{d}}_{t+5}$, $\hat{\mathbf{d}}_{t+10}$, between the ground truth and prediction frames.

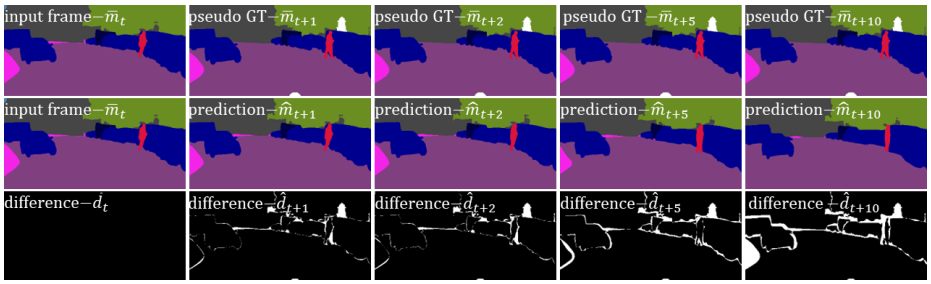

Fig. 2: Output predictions for a sequence of the Cityscapes validation split, $\mathcal{D}_{\text{val}}^{\text{CS-vid}}$. The top row depicts actual 20^{th} frame ground truth annotations available in the dataset for $\overline{\mathbf{m}}_t$, $\overline{\mathbf{m}}_{t+1}$, $\overline{\mathbf{m}}_{t+2}$, $\overline{\mathbf{m}}_{t+5}$, $\overline{\mathbf{m}}_{t+10}$. In the middle row, we show the input semantic segmentation $\overline{\mathbf{m}}_t$ along with the predictions $\hat{\mathbf{m}}_{t+1}$, $\hat{\mathbf{m}}_{t+2}$, $\hat{\mathbf{m}}_{t+5}$, $\hat{\mathbf{m}}_{t+10}$ from the prediction network. The bottom row portrays the absolute difference $\hat{\mathbf{d}}_{t+1}$, $\hat{\mathbf{d}}_{t+2}$, $\hat{\mathbf{d}}_{t+5}$, $\hat{\mathbf{d}}_{t+10}$, between the ground truth and prediction frames.

1 Qualitative Results

In this section, we show the qualitative results of our method on sequences of BDD100K, $\mathcal{D}_{\text{val}}^{\text{BDD-MOTS}}$ and Cityscapes, $\mathcal{D}_{\text{val}}^{\text{CS-vid}}$. In Figure 1, we show the qualitative results of prediction method on a sequence of $\mathcal{D}_{\text{val}}^{\text{BDD-MOTS}}$. We can observe that, for increasing time steps, i.e., $\Delta t = \{1, 2, 5, 10\}$, the prediction

Table 1: Re-ordering of semantic classes in BDD100K

class index (s)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
original order	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	veget -ation	terrain	sky	person	rider	car	truck	bus	train	motor -cycle	bicycle
first reorder	pole	sky	wall	traffic sign	road	truck	train	fence	sidewalk	bicycle	person	traffic light	veget -ation	building	bus	car	rider	terrain	motor -cycle
second reorder	train	bicycle	bus	terrain	rider	sky	sidewalk	road	wall	traffic sign	motor -cycle	traffic light	person	fence	building	truck	car	pole	veget -ation

Fig. 3: A semantic segmentation input mask $\overline{\mathbf{m}}_t$ from $\mathcal{D}_{\text{val}}^{\text{BDD-MOTS}}$ showing different semantic classes. The left white encircled region portrays the proximal occurence of class `sidewalk` ($s = 2$) near to class `road` ($s = 1$). Similarly, the right white encircled region portrays the proximal occurrence of class `road` ($s = 1$) near to class `car` ($s = 14$).

worsens for dynamic objects. This can be inferred from the increase in white regions in the absolute difference estimation visualizations (bottom row) defined as $\hat{\mathbf{d}}_t = |\hat{\mathbf{m}}_t - \overline{\mathbf{m}}_t|$. Also, majority of the predictions are incorrect in the boundary segregation of different classes, i.e., where car pixel occupancy meets road occupancy, where sidewalk occupancy meets road occupancy.

Specifically for the Cityscapes dataset, in Figure 2, the predictions (middle row) are at par with their corresponding 20^{th} frame ground truth annotations (top row) for $\Delta t = \{1, 2, 5\}$, obtained from the dataset. The semantic class boundaries are so well captured with noise supression in predictions (middle row). However, for $\Delta t = 10$, we can see that the prediction focuses more on predicting the static classes than the finer dynamic classes boundary details, i.e., visible from the missing sidewalk in the left region of $\hat{\mathbf{m}}_{t+10}$ that is visible in the ground truth annotation $\overline{\mathbf{m}}_{t+10}$ (in pink) of Figure 2.

2 Prediction Invariance on the Ordering of Semantic Classes

We investigate the behavior of our model when 1-channel inputs are fed to our predictor network, i.e., the generated pseudo ground truth $\overline{\mathbf{m}}_t \in \mathcal{S}^{H \times W \times 1}$ for the BDD100K dataset $\mathcal{D}^{\text{BDD-MOTS}}$. The semantic segmentation mask $\overline{\mathbf{m}}_t$ contains class indices $s \in \mathcal{S} = \{1, 2, ..., S\}$, where $S = 19$. Here, each semantic class corresponds to a specific class index s, e.g., $s = 0$ for class `road`, $s = 12$ for

Table 2: Confusion matrix on BDD100K: scores of all the $S = 19$ classes in BDD100K with original ordering. The classes with highest true score are highlighted in red and the cells with second highest true score are marked in light red.

Predicted Class →

Ground Truth ↓

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
road	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
sidewalk	0.14	0.76	0.02	0.01	0.01	0.01	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
building	0.00	0.00	0.90	0.00	0.01	0.01	0.00	0.00	0.03	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
wall	0.04	0.02	0.12	0.65	0.07	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
fence	0.02	0.02	0.1	0.04	0.71	0.01	0.00	0.00	0.05	0.01	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
pole	0.03	0.02	0.17	0.01	0.01	0.53	0.00	0.01	0.1	0.01	0.08	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
traffic light	0.01	0.00	0.21	0.00	0.00	0.05	0.57	0.03	0.05	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
traffic sign	0.00	0.00	0.13	0.00	0.01	0.02	0.00	0.75	0.05	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
vegetation	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.9	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
terrain	0.12	0.09	0.02	0.01	0.01	0.01	0.00	0.00	0.08	0.64	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
sky	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
person	0.12	0.04	0.1	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.64	0.01	0.05	0.00	0.00	0.00	0.00	0.00
rider	0.08	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.49	0.14	0.1	0.01	0.01	0.00	0.03	0.02
car	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00
truck	0.04	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.1	0.75	0.01	0.00	0.00	0.01
bus	0.03	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.05	0.04	0.78	0.00	0.00	0.00
train	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
motorcycle	0.33	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.05	0.13	0.02	0.02	0.00	0.15	0.15
bicycle	0.25	0.02	0.03	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.07	0.00	0.01	0.00	0.01	0.58

class `person`, etc according to the scene. For instance, if we consider a small region in a semantic segmentation mask, we usually find a certain semantic class pixels in close proximity with another semantic class pixels, i.e, class `road` pixels ($s = 1$) almost always occurs adjacent to class `car` pixels ($s = 14$) and class `sidewalk` pixels ($s = 2$) almost are always adjacent to class `road` pixels ($s = 1$) as can be seen in Figure 3. To investigate the proposed method's performance and robustness when the original class orientation is re-ordered, we conducted some experiments by shuffling the class indexes in the generated pseudo ground truth frames $\overline{\mathbf{m}}_t$. For instance, now the same scene would contain class `road` ($s = 5$) adjacent to class `car` ($s = 16$) and class `sidewalk` ($s = 9$) adjacent to class `road` ($s = 5$). Note that, the semantic classes remain the same, just the class indices are shuffled randomly. In Table 1, we can see the original class order along with the re-ordered class indices where semantic classes are marked by their actual defined colors in $\mathcal{D}^{BDD-MOTS}$. can This is an important investigation to

Table 3: Confusion matrix on BDD100K: scores of all the $S = 19$ classes in BDD100K with first re-ordering. The classes with highest true score are highlighted in red and the cells with second highest true score are marked in light red.

Predicted Class

Ground Truth

	pole	sky	fence	traffic sign	road	truck	train	fence	sidewalk	bicycle	person	traffic light	vegetation	building	bus	car	rider	terrain	motorcycle
pole	0.54	0.09	0.01	0.01	0.03	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.08	0.17	0.00	0.01	0.00	0.01	0.00
sky	0.00	0.92	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.00	0.00
wall	0.01	0.01	0.67	0.00	0.04	0.00	0.00	0.07	0.02	0.00	0.00	0.00	0.04	0.11	0.00	0.02	0.00	0.01	0.00
traffic sign	0.01	0.04	0.00	0.75	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.13	0.00	0.00	0.00	0.00	0.00
road	0.00	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
truck	0.00	0.01	0.00	0.01	0.06	0.78	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.06	0.01	0.05	0.00	0.00	0.00
train	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fence	0.01	0.01	0.05	0.00	0.02	0.00	0.00	0.73	0.02	0.00	0.00	0.00	0.04	0.09	0.00	0.02	0.00	0.01	0.00
sidewalk	0.00	0.00	0.01	0.00	0.15	0.00	0.00	0.01	0.75	0.00	0.01	0.00	0.01	0.02	0.00	0.01	0.00	0.03	0.00
bicycle	0.01	0.00	0.00	0.00	0.28	0.02	0.00	0.01	0.04	0.55	0.02	0.00	0.01	0.02	0.00	0.03	0.00	0.00	0.01
person	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.01	0.04	0.01	0.63	0.00	0.01	0.1	0.00	0.04	0.02	0.00	0.00
traffic light	0.02	0.08	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.06	0.2	0.00	0.00	0.00	0.00	0.00
vegetation	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.89	0.05	0.00	0.01	0.00	0.01	0.00
building	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.90	0.00	0.01	0.00	0.00	0.00
bus	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.09	0.76	0.05	0.00	0.00	0.00
car	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.92	0.00	0.00	0.00
rider	0.00	0.00	0.00	0.00	0.1	0.01	0.00	0.00	0.01	0.01	0.34	0.00	0.02	0.1	0.00	0.11	0.24	0.02	0.02
terrain	0.00	0.00	0.01	0.00	0.13	0.00	0.00	0.01	0.07	0.00	0.00	0.00	0.07	0.02	0.00	0.02	0.00	0.66	0.00
motorcycle	0.00	0.00	0.00	0.00	0.38	0.01	0.00	0.00	0.03	0.05	0.12	0.00	0.00	0.01	0.00	0.17	0.11	0.03	0.1

prove that our predictor model still learns the proximal relationship between the semantic classes instead of the numerical class indices occupancy, i.e., our model perfectly learns that class `road` pixels are most likely to occur near class `car` pixels and vice-versa, irrespective of their class indices value. Hence, we performed experiments by re-ordering the class indices of $\mathcal{D}^{\text{BDD-MOTS}}$ in such a way that the classes that occurred near to each other in terms of class index distance, e.g., `road` ($s = 1$) and `sidewalk` ($s = 2$) are now placed further apart, e.g., `road` ($s = 5$) and `sidewalk` ($s = 9$) as can be seen in Table 1.

Table 2 shows the confusion matrix for the original class order of $\mathcal{D}^{\text{BDD-MOTS}}$. The confusion matrix represents how each class in the prediction is confused and interpreted with respect to all the classes present in the ground truth and vice-versa. It can be observed that, every class is predicted well with the highest score for itself (see diagonal) except class `rider` ($s = 13$) which is predicted as class `person` ($s = 12$) with a score of 0.49 which is obvious as `rider` fits into

Table 4: Confusion matrix on BDD100K: scores of all the $S = 19$ classes in BDD100K with second re-ordering. The classes with highest true score are highlighted in red and the cells with second highest true score are marked in light red.

Predicted Class →

Ground Truth ↓

	train	bicycle	bus	terrain	rider	sky	sidewalk	road	wall	traffic sign	motorcycle	traffic light	person	fence	building	truck	car	pole	vegetation
train	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bicycle	0.00	0.50	0.04	0.00	0.00	0.00	0.02	0.29	0.00	0.00	0.01	0.00	0.01	0.01	0.03	0.04	0.04	0.00	0.01
bus	0.00	0.01	0.76	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.1	0.03	0.05	0.00	0.01
terrain	0.00	0.00	0.00	0.63	0.00	0.01	0.09	0.14	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.08
rider	0.00	0.03	0.02	0.01	0.13	0.01	0.02	0.12	0.00	0.00	0.05	0.00	0.37	0.00	0.13	0.01	0.07	0.00	0.02
sky	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.03
sidewalk	0.00	0.00	0.00	0.02	0.00	0.01	0.75	0.15	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.01	0.00	0.01
road	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
wall	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.04	0.66	0.00	0.00	0.00	0.00	0.07	0.12	0.00	0.02	0.00	0.05
traffic sign	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.01	0.05
motorcycle	0.00	0.08	0.01	0.00	0.04	0.00	0.02	0.38	0.00	0.01	0.17	0.00	0.16	0.00	0.01	0.01	0.11	0.00	0.00
traffic light	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.01	0.00	0.02	0.00	0.63	0.00	0.00	0.2	0.00	0.00	0.02	0.05
person	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.15	0.00	0.00	0.00	0.00	0.63	0.01	0.1	0.00	0.04	0.00	0.01
fence	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.02	0.04	0.00	0.00	0.00	0.00	0.71	0.11	0.00	0.02	0.01	0.04
building	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.91	0.00	0.01	0.01	0.03
truck	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.08	0.75	0.07	0.00	0.02
car	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.92	0.00	0.01
pole	0.00	0.00	0.00	0.01	0.00	0.08	0.02	0.03	0.01	0.01	0.00	0.00	0.00	0.01	0.17	0.00	0.02	0.5	0.12
vegetation	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.00	0.91

the broader category of `person` after all. Similarly, class `motorcycle` ($s = 18$) gets confused for class `road` ($s = 1$) with a score of 0.33. This could be simply attributed the fact that the class road heavily overpowers the pixel distribution in all scenes whereas the class motorcycle has very minimal occupancy in most of the scenes. Now, in Table 3, we can see the confusion matrix for the first re-ordering of classes. In Table 3, it can be observed that, the predictor still confuses class `rider` ($s = 17$) for class `person`($s = 11$) with a score of 0.34 and class `motorcycle` ($s = 19$) for class `road` ($s = 5$) with a score of 0.38. Similarly, in Table 4, we can see the confusion matrix for the second re-ordering of classes. We can see that the predictor once again interprets class `rider` ($s = 5$) as class `person` ($s = 13$) with a score of 0.37 and class `motorcycle` ($s = 11$) as class `road` ($s = 8$) with a score of 0.38. It can be inferred that the predictor regardless of class index ordering, behaves exactly the same for all the class predictions. Thus, the predictor can be safely labeled as invariant to the class ordering.