Guiding Attention in End-to-End Driving Models

Diego Porres [1] Yi Xiao [1] Gabriel Villalonga [1] Alexandre Levy [1] Antonio M. López [1 2]

[1]Computer Vision Center [2]Universitat Autònoma de Barcelona

Website Paper Code

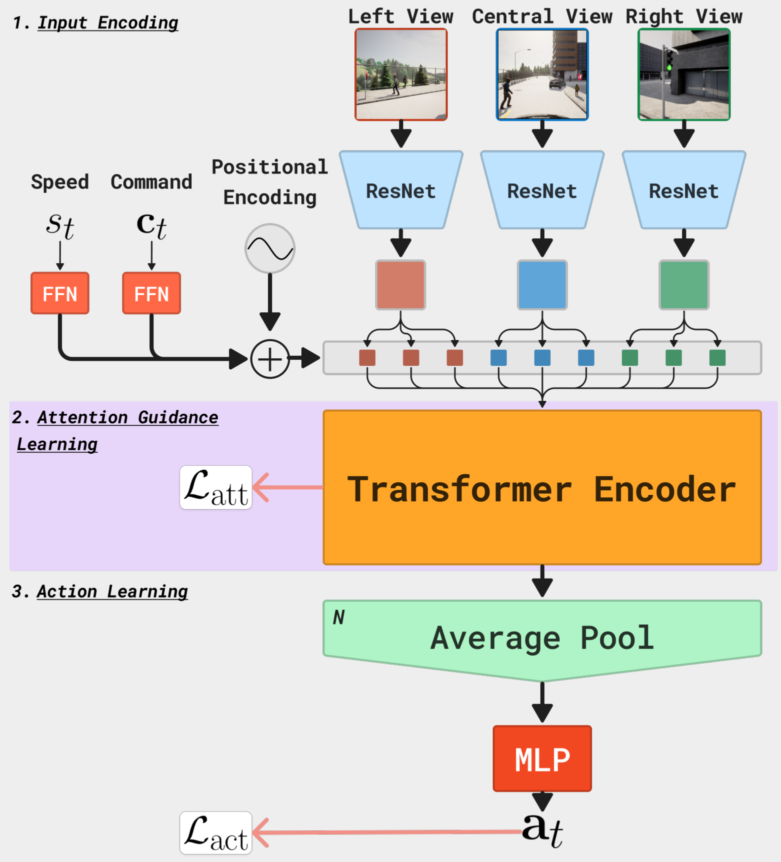
Problem Formulation & Motivation

With Imitation Learning (IL), driving policies seek to **approximate the driving behavior of the expert driver** that collects the training data. Vision-based end-to-end driving trained via IL offer **affordable solutions** for autonomous driving, albeit they require **large amounts of data** in order to properly converge.

In this paper, we study the effects of **directly optimizing the attention maps** on the driving capabilities of these models and their interpretability. We show that the model's **sample efficiency improves**, highlighted when there is a low amount of data to train with.

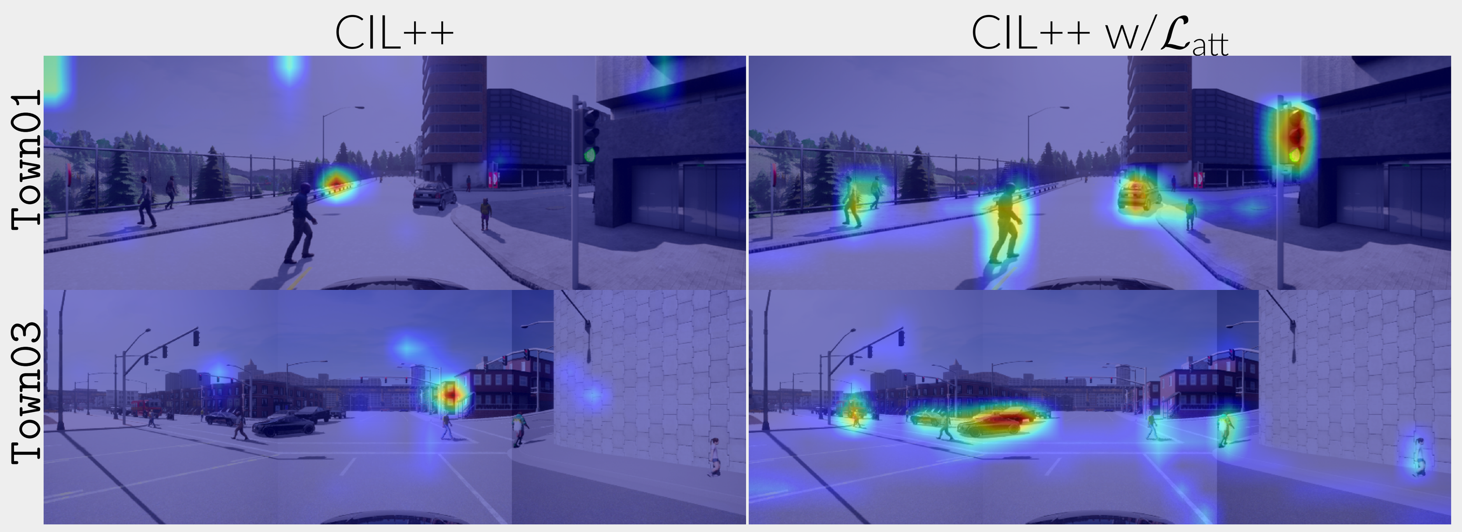
What if we directly optimize the self-attention weights?

We base our work on the current pure vision-based state-of-the-art end-t-end driving model **CIL++**.



Benefits

- Adding the Attention Loss $\mathcal{L}_{att}$ during training **circumvents the need to predict the attention masks during validation**, nor to modify the original architecture.
- The model's interpretability is improved, as the **attention weights now weakly segment the classes of interest** (pedestrian, vehicles, traffic lights, lane lines, and curb).
- The model also **needs less data** to get the same driving quality compared to the vanilla CIL++, and is **robust to noisy attention masks**.



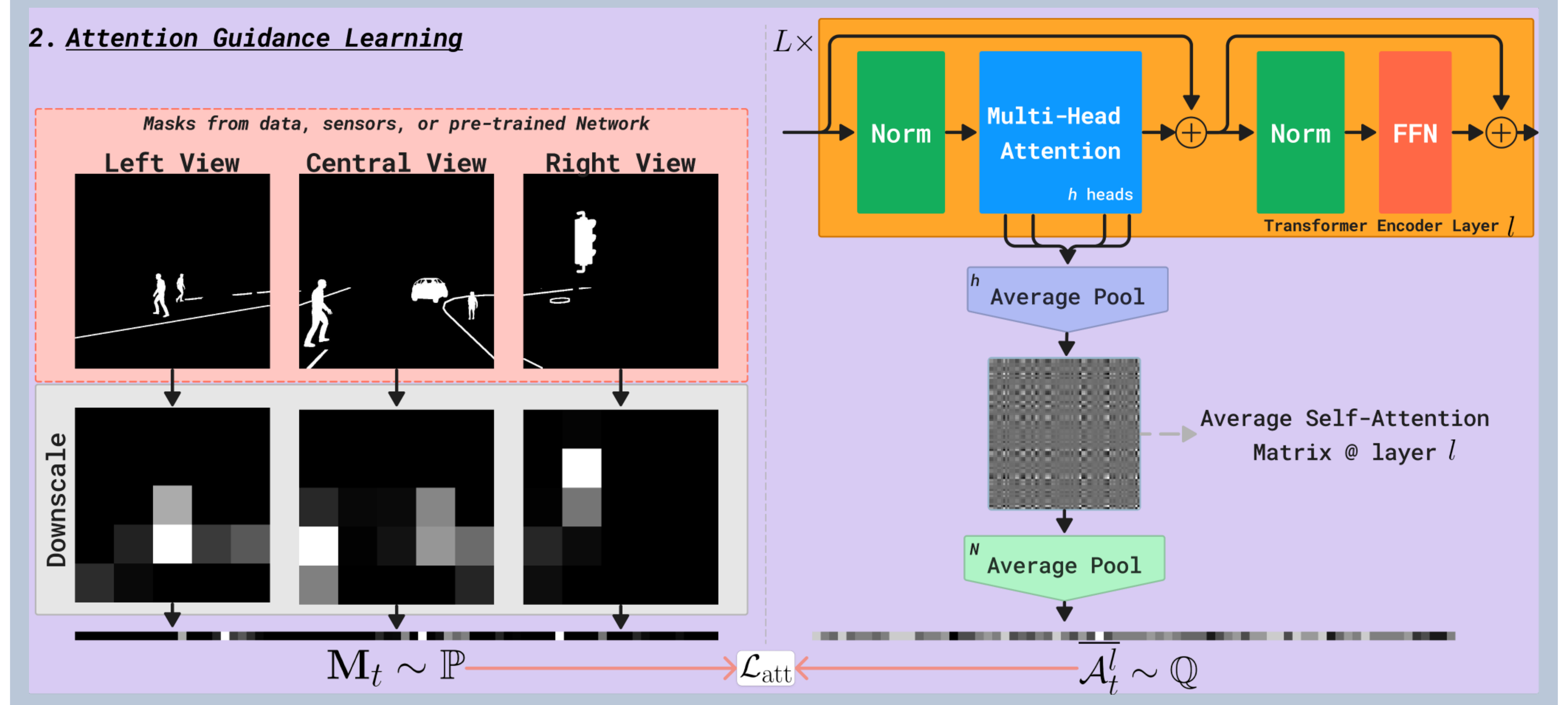
Future Work

$\mathcal{L}_{att}$ could be applied not only to the average attention weight of a layer in the Transformer Encoder, but to their **individual heads**. Likewise, the attention masks could also come from **human saliency maps** collected during driving.

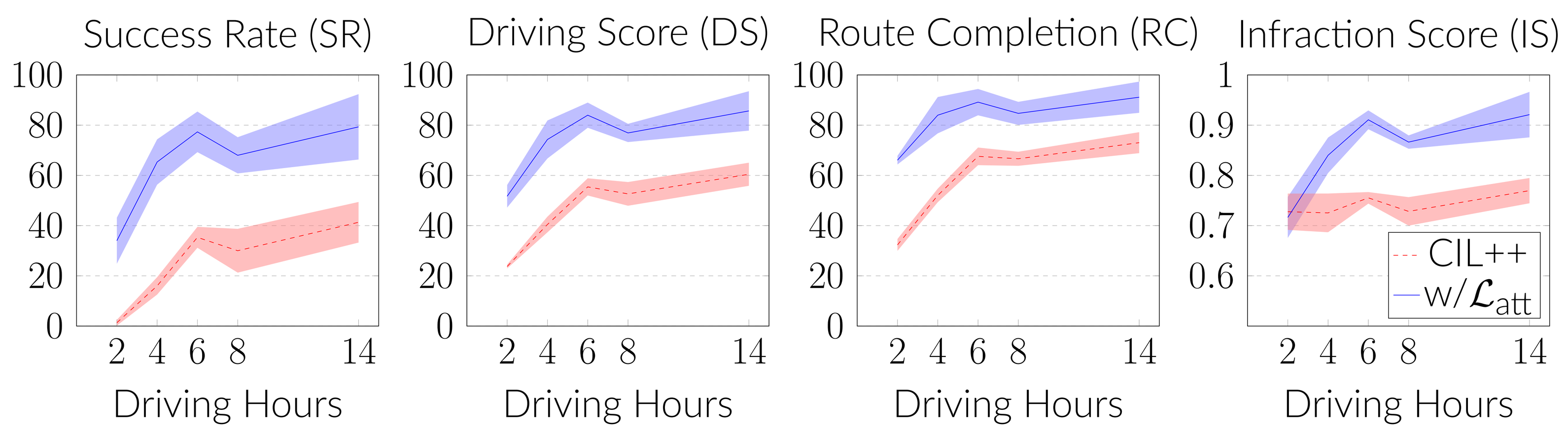
Our proposal: the Attention Loss $\mathcal{L}_{att}$

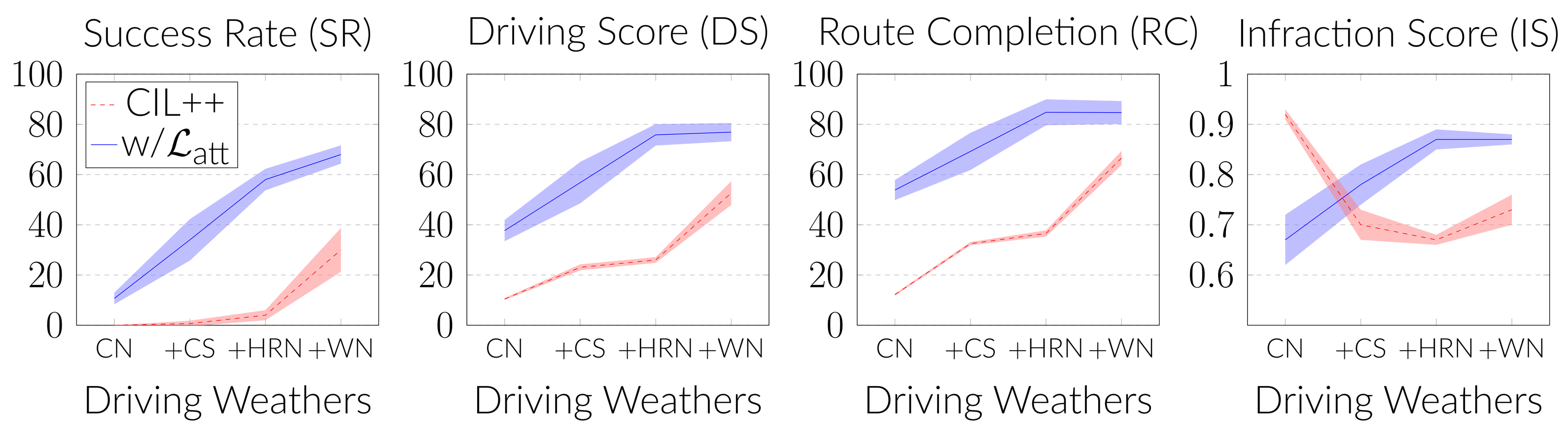
We wish to exploit the **distributional property** of the attention weights of the Transformer Encoder. For this, we create ground-truth single-channel *synthetic* attention masks $\mathcal{M}_{i,t}$ for each camera i based on Semantic Segmentation images (containing the classes of interest), filtered within a depth threshold.

We define the **Attention Loss** $\mathcal{L}_{att}$ as the KL Divergence between the (downscaled, concatenated, and normalized) ground-truth saliency maps $\mathbf{M}_t$ and the average attention weights of layer l of the Transformer Encoder $\mathcal{A}_t^l$ at time t.



~ 4× less training data for the same driving capability!





$\mathcal{L}_{att}$ is robust to noisy saliency masks

Obtaining the *synthetic attention* masks for real-world data will result in **noisy masks**. We mimic this noise via a function f that corrupts the mask $\mathcal{M}_{i,t}$ using depth-aware Perlin noise, with more granular disturbances on larger objects. As a proxy, we train a UNet to predict the mask $\widehat{\mathcal{M}}_{i,t}$ given an input image $\mathbf{x}_{i,t}$.

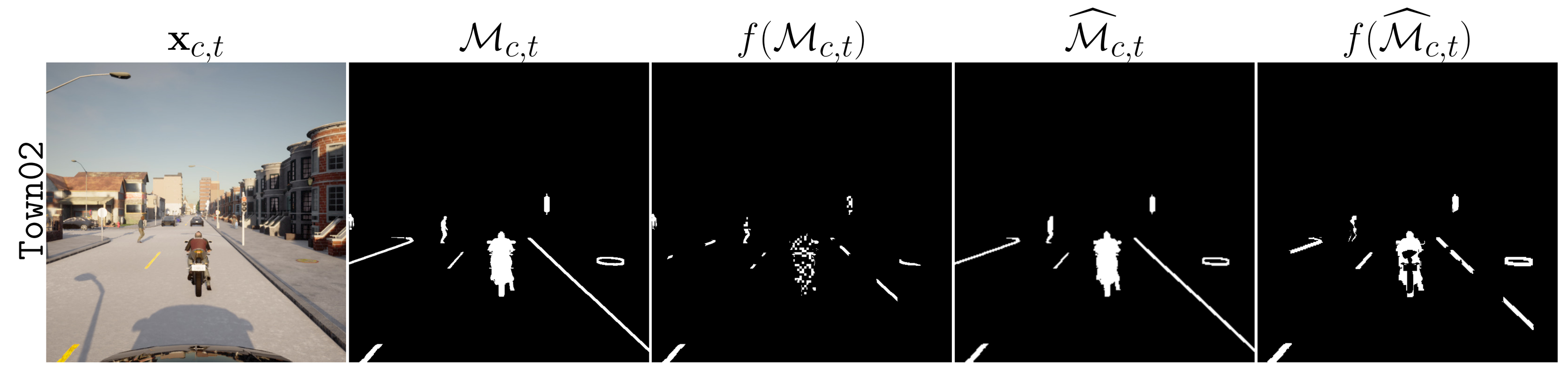


Table 1. Masks as different types of input and effect of noisy masks. Models trained with 14 hours of data from `Town01` and tested in `Town02`, using new weathers.

	SR ↑	DS ↑	RC ↑	IS ↑
CIL++	41.33 ± 8.08	60.45 ± 4.60	73.03 ± 4.18	0.77 ± 0.03
w/SM	42.00 ± 7.21	59.29 ± 5.49	70.12 ± 4.32	0.78 ± 0.02
w/HM	66.00 ± 9.17	77.34 ± 6.93	84.32 ± 5.83	0.87 ± 0.04
w/$\mathcal{L}_{att}$	$\mathbf{79.33} \pm 13.01$	$\mathbf{85.67} \pm 7.84$	$\mathbf{91.13} \pm 6.21$	$\mathbf{0.92} \pm 0.05$
w/SM + $f(\widehat{\mathcal{M}}_{i,t})$ [a]	35.33 ± 7.02	56.38 ± 1.32	68.38 ± 0.58	0.77 ± 0.01
w/HM + $f(\widehat{\mathcal{M}}_{i,t})$	66.00 ± 7.21	76.36 ± 3.72	83.46 ± 4.48	0.87 ± 0.01
w/$\mathcal{L}_{att}$ + $f(\mathcal{M}_{i,t})$ [b]	$\mathbf{71.33} \pm 6.11$	$\mathbf{80.36} \pm 6.88$	$\mathbf{89.46} \pm 3.97$	$\mathbf{0.87} \pm 0.05$

[a] Noisy predicted Masks (Training + Validation) [b] Noisy Masks (Training only)

Table 2. Effect of using $\mathcal{L}_{att}$ in the high-data regime for multi-lane towns in CARLA. Models trained with 55 hours of driving data and tested in the unseen `Town05`, using new weathers.

	SR ↑	DS ↑	RC ↑	IS ↑
CIL++	70.00 ± 5.00	36.46 ± 4.03	79.69 ± 3.84	0.51 ± 0.04
w/$\mathcal{L}_{att}$	$\mathbf{73.33} \pm 5.77$	$\mathbf{58.23} \pm 4.71$	$\mathbf{82.88} \pm 1.28$	$\mathbf{0.70} \pm 0.03$

Acknowledgements

This research is supported by project TED2021-132802BI00 funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. Antonio M. Lopez acknowledges the financial support to his general research activities given by ICREA under the ICREA Academia Program. Antonio and Gabriel thank the synergies, in terms of research ideas, arising from the project PID2020-115734RB-C21 funded by MCIN/AEI/10.13039/501100011033. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program and its ACCIO agency to CVC's general activities.