On the Road to Portability: Compressing End-to-End Motion Planner for Autonomous Driving

Supplementary Material

Appendix

A. Implementation Details

A.1. Hyper-parameter Setting

We perform all the experiments using GeForce RTX 3090 GPU. As for training, we use the Adam optimizer [2] for optimization with a learning rate 0.005 for InterFuser and 0.0001 for TCP. For TCP backbones, the epoch number is set to 30 and the batch size is set to 24. For InterFuser backbones, the epoch number is set to 10 and the batch size is set to 16. We empirically set $\alpha_r = 0.1$. α_z is set to 0.1 for InterFuser and 0.5 for TCP. α_e is set to 0.005 for InterFuser and 0.05 for TCP. The standard deviation σ in the kernel function for adjusting the smoothness is set to 3. The action threshold δ is set to 0.1. The Lagrange multiplier β is set to 0.001. Following the original papers [5, 6], we set the number of planned waypoints to $T = 4$ for TCP and $T = 10$ for InterFuser.

A.2. Image Resolution Setting

For InterFuser, the resolution of camera image is 224×224 for the front camera and 128×128 for the side camera, following the original paper [5]. For TCP, the resolution of camera image is 900×256 for the front camera, following the original paper [6]. The horizontal field of view for all cameras is set as 100°. In our method, the BEV scene image is provided by the CARLA simulator [1] and is with a resolution of 180×180.

A.3. Lightweight Planner Architecture Setting

As aforementioned in the main body of the paper, the end to end motion planners can generally be divided into two main parts: the perception backbone and the motion producer. In this paper, we reduce the number of the parameters of these two parts by taking smaller backbones as lightweight planners. The detailed configurations of the original planners and the corresponding lightweight planners are presented in Table 1.

A.4. Planning-relevant Feature Distillation Module Setting

In the planning-relevant feature distillation module, we configure both the encoder and decoder of the information bottleneck by using a 3-layer MLPs with a hidden size of 512 and LeakyReLU [3] as the activation function. The dimension of the planning-relevant feature is set to 256. To transfer planning-relevant knowledge, We empirically select the middlemost layer of the teacher's perception backbone to distill the knowledge to the middlemost layer of the the student's perception backbone. Before inputting the intermediate feature map to the information bottleneck encoder, we perform channel-wise averaging along the channel dimension.

A.5. Safety-aware Waypoint-attentive Distillation Module Setting

In the safety-aware waypoint-attentive distillation module, the BEV encoder consists of a 6-layer CNN followed by a 2-layer MLP with a hidden size of 512. The waypoint encoder, on the other hand, is configured as a 2-layer MLP with a hidden size of 128. Both of these two encoders utilize the LeakyReLU activation function. For simplicity, we adopt the expert waypoints as the teacher waypoints in this paper. In the attention mechanism, both of the dimensions of the query and the key are set to 64.

B. Training Pseudo-code

The training procedure for our PlanKD method is outlined in Algorithm 1. Firstly, the student planner and teacher planner undergo forward propagation to obtain their intermediate feature maps and output waypoints. These intermediate feature maps are then passed through the information bottleneck encoder to extract the planning-relevant features. Using the planning-relevant features from both the teacher planner and the student planner, we calculate the planning-relevant knowledge distillation loss. Next, the planning-relevant features are input to the information bottleneck decoder to obtain the planning states, which are used to compute the upper bound of the information bottleneck objective. Moving on to the safety-aware waypoint-attentive distillation module, we determine the importance of the teacher's waypoints and the expert's waypoints. Based on the obtained importance weights, we calculate the safety-aware waypoint loss, as well as the safety-aware ranking loss and the entropy loss. Finally, all these losses are aggregated and used as the overall loss for optimization. By employing PlanKD during the training process, we can develop a portable and safe planner suitable for deployment in resource-limited environments.

Table 1. Configurations of different planners. Transformer-3 (128) denotes a 3-layer transformer with an embedding size of 128. MLPs-half denotes MLPs with half of the original hidden size. The inference time per frame is evaluated on GeForce RTX 3090 GPU.

Backbone	Parameter Count	Camera Perception Backbone	LiDAR Perception Backbone	Motion Producer Backbone	Model FLOPS	Inference Time (ms)
InterFuser	52.9M	ResNet-50	ResNet-18	Transformer-6 (256)	46.51G	78.3
	26.3M	ResNet-18	ResNet-18	Transformer-3 (128)	25.52G	39.7
	11.7M	ResNet-10	ResNet-10	Transformer-3 (64)	11.12G	22.8
	3.8M	ResNet-6	ResNet-6	Transformer-2 (64)	7.21G	17.2
TCP	25.8M	ResNet-34	-	MLPs	17.09G	17.9
	13.9M	ResNet-18	-	MLPs-half	8.47G	10.7
	7.6M	ResNet-10	-	MLPs-half	4.15G	8.5
	3.1M	ResNet-6	-	MLPs-half	2.67G	7.2

Algorithm 1 The training procedure of PlanKD

Input: a pretrained large teacher planner $\mathcal{F}_\theta^T$, dataset $\mathcal{D} = \{(I, \mathcal{T}^*)\}$, ground truth planning states Y^i, BEV scene representation B, epochs N_e;
Output: a trained compact student planner $\mathcal{F}_\phi^S$;
Initialize the parameters of student planner $\mathcal{F}_\phi^S$;
Initialize the parameters of the two modules in PlanKD;
Freeze the parameters of teacher planner $\mathcal{F}_\theta^T$;
for each epoch e from 1 to N_e **do**
  for each batch b in epoch e **do**
    obtain the intermediate feature map h^T of teacher planner $\mathcal{F}_\theta^T$;
    obtain the intermediate feature map h^S of student planner $\mathcal{F}_\phi^S$;
    input h^T,h^S to IB encoder to derive planning-relevant feature z^T,z^S;
    calculate the planning-relevant knowledge distillation loss $\mathcal{L}_z$;
    input z^T,z^S to IB decoder to derive the prediction of planning states;
    calculate the upper bound of the information bottleneck objective $\mathcal{L}_{IB}$;
    derive the attention weight between teacher waypoint w_i^T and B;
    derive the attention weight between expert waypoint w_i^* and B;
    calculate the safety-aware waypoint loss $\mathcal{L}_w$ and $\mathcal{L}_{w*}$;
    calculate the ranking loss $\mathcal{L}_{rank}$ the entropy loss $\mathcal{L}_e$;
    calculate the overall loss $\mathcal{L}$;
    optimize the learnable parameters by $\mathcal{L}$;
  end for
end for

C. Additional Experiments

C.1. Additional Comparison with KD

Here, we present additional comparison results with other knowledge distillation methods on the Town05 Long Benchmark. As shown in Table 2, it is evident that our PlanKD method continues to outperform previous knowledge distillation methods by a significant margin.

C.2. Additional Ablation Study

To further validate the effectiveness of our method, we perform an ablation study on the Town05 Long Benchmark, as presented in Table 3. The results further demonstrate the effectiveness of each component in our proposed method.

C.3. Additional Visualizations

To investigate the planning-relevant knowledge extracted by the information bottleneck, we employ the Grad-CAM technique [4] to visualize the intermediate feature maps of InterFuser. The visualization is guided by the gradient of the planning states within the information bottleneck, revealing where the extracted planning-relevant knowledge is concentrated. The results are presented in Figure 1. Figure 1(a) represents a normal scene with no moving obstacles. The planning-relevant knowledge focuses on the lanes, in-

Table 2. Comparisons with other knowledge distillation methods on the Town05 Long Benchmark.

Method	Backbone	Teacher Param	Student Param	Driving Score(↑)	Route Completion(↑)	Infraction Score(↑)	Collision Rate(↓)	Infraction Rate(↓)
AT	InterFuser	52.9M	26.3M	41.62	85.61	0.472	0.112	0.134
ReviewKD		52.9M	26.3M	40.67	93.25	0.426	0.178	0.168
DPK		52.9M	26.3M	44.29	81.10	0.550	0.095	0.113
PlanKD (Ours)		52.9M	26.3M	55.90	97.44	0.562	0.094	0.093
AT	TCP	25.8M	13.9M	43.31	100.0	0.433	0.159	0.128
ReviewKD		25.8M	13.9M	41.27	94.64	0.431	0.148	0.147
DPK		25.8M	13.9M	43.83	90.27	0.499	0.158	0.146
PlanKD (Ours)		25.8M	13.9M	53.19	93.28	0.579	0.084	0.116

Table 3. Ablation Study of PlanKD on the Town05 Long Benchmark.

Method	Backbone	Teacher Param	Student Param	Driving Score(↑)	Route Completion(↑)	Infraction Score(↑)	Collision Rate(↓)	Infraction Rate(↓)
PlanKD-w.o.-entropy	InterFuser	52.9M	26.3M	46.73	70.49	0.643	0.141	0.063
PlanKD-w.o.-safe-att		52.9M	26.3M	44.55	75.37	0.555	0.141	0.097
PlanKD-w.o.-IB		52.9M	26.3M	50.17	92.72	0.509	0.162	0.111
PlanKD		52.9M	26.3M	55.90	97.44	0.562	0.094	0.093
PlanKD-w.o.-entropy	TCP	25.8M	13.9M	45.72	71.64	0.668	0.088	0.127
PlanKD-w.o.-safe-att		25.8M	13.9M	45.07	100.0	0.450	0.160	0.121
PlanKD-w.o.-IB		25.8M	13.9M	50.70	100.0	0.507	0.096	0.130
PlanKD		25.8M	13.9M	53.19	93.28	0.579	0.084	0.116

dicating the importance of keeping lane for the ego-vehicle. In Figure 1(b), where a pedestrian suddenly appears, the planning-relevant knowledge is directed towards the pedestrian, highlighting the need to avoid collision. Figure 1(c) showcases a situation where a vehicle is in front and a motorbike is driving towards the ego-vehicle. In this case, it's important to maintain a safe distance, thus the planning-relevant knowledge emphasizes other road users. Figure 1(d) depicts a scenario with a traffic light, where the attention is drawn to the state of the traffic light. Finally, Figure 1(e) shows an intersection scenario where the ego-vehicle requires extra caution to interact with other road users. Thus, the planning-relevant knowledge focuses on the interacting vehicle in front. These visualizations indicate that our method can successfully extracts the knowledge that are significant to planning across various scenarios.

Besides, we also visualize the attention maps generated by the knowledge distillation method AT [7]. It can be observed that the generated attention maps contain numerous planning-irrelevant information (especially in Figure 1(a)(b)(c)). This further indicates the superiority of our method.

D. Limitations and Future Works

Our work mainly focus on the knowledge distillation technique for compressing end-to-end motion planner in autonomous driving. Exploring the integration of other model compression techniques, such as quantization and pruning, into our approach is a promising avenue for future research. By doing so, we can further reduce the size of the motion planner and enhance its efficiency.

Besides, we devise a simple yet effective way to take the safety significance of each waypoint into account via the learning-based attention. In the future, it is possible to incorporate specific expert knowledge about driving to design a more comprehensive and refined strategy for determining the importance of waypoints. In addition, the current method primarily emphasizes the proximity of waypoints to obstacles as a measure of danger, which captures an important aspect of safety. The approach is grounded in the fact that immediate physical distance from obstacles is a critical factor in potential collisions. While our current approach prioritizes spatial proximity to obstacles, incorporating temporal aspects, could indeed offer a more comprehensive safety assessment.

Furthermore, in our approach, knowledge transfer in

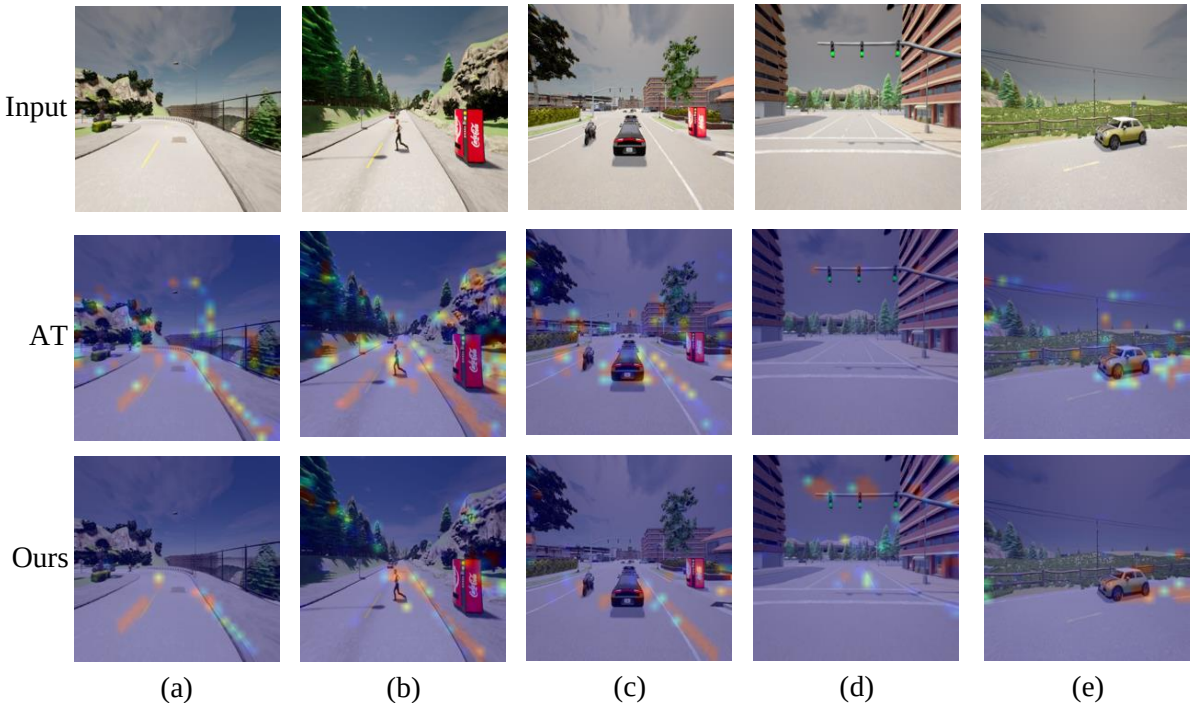

Figure 1. Visualizations of the intermediate feature maps of InterFuser. The redder regions represent higher activation values. The first row is the input image of the front camera. The second row is the corresponding attention map generated by AT [7]. The third row is the corresponding Grad-CAM [4] visualization guided by the gradient of the planning states in the information bottleneck.

the intermediate layer is currently limited to feature maps within the same sensor modality. For planners that incorporate multiple sensor modalities, a potential future direction could involve developing methods to distill knowledge between different sensors to facilitate cross-modal knowledge transfer.

Finally, our method trained on CARLA is subject to the well-known simulation-to-reality gap, which implies that its performance might differ when deployed in the real world. This necessitates extensive real-world testing and validation to ensure that the model's behavior aligns with expected safety norms. Safety assurance processes must encompass a wide range of scenarios and edge cases that vehicles might encounter, ensuring the model's robustness and reliability.

References

[1] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, page 3. Atlanta, Georgia, USA, 2013. 1

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2, 4

[5] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 1

[6] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *Advances in Neural Information Processing Systems*, 2022. 1

[7] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. 3, 4