1. Experimental Setup

1.1. Implementation Details

Here we further provide detailed experimental settings in our paper. Class numbers for object detection, semantic segmentation, drivable area segmentation, and lane detection are 9, 19, 2, and 1 respectively. We remove the train class as in [12] for object detection. For lane detection, we follow [5] to preprocess lane line annotations. Loss weights for object detection, semantic segmentation, drivable area segmentation, and lane detection are fixed as 1, 2, 2, and 2 respectively. All experiments are conducted on servers with 8 Nvidia V100 GPUs and Intel Xeon Platinum 8168 CPU (2.70GHz).

1.2. More Details on Dataset

BDD100k dataset [14] contains multiple tasks. Here we focus on object detection (OD), semantic segmentation (SS), drivable area segmentation (DA), and lane detection (LD). In BDD100K, 70k training images are labeled for object detection, drivable area segmentation, and lane detection, and only 7k training images are labeled for semantic segmentation.

2. More Investigations

In this section, we present more analyses of popular multi-task learning methods.

2.1. Task Scheduling

Here we analyze task scheduling methods on disjoint-balance settings, whose results are shown in Table 1 in the main paper. Note that the full setting contains almost complete annotations except semantic segmentation, thus it is not suitable for task scheduling. Since the data of all tasks in the disjoint-balance setting is balanced and non-overlapped, Uniform sampler [9] and Weighted sampler [9] are equivalent. As shown in Table 1 in the main paper, task sampling methods (i.e., Uniform sampler [9] and Round-robin [9]) perform better than Zeroing loss [13] by a large margin on segmentation-based tasks, but get worse in object detection. We hypothesize that negative transfer still exists among these approaches, and training one task per step may lead to forgetting to some extent.

2.2. Partial-label Learning

As shown in Table 1 in the main paper, pseudo labeling [3] surpasses Zeroing loss [13] on almost all tasks, especially on semantic segmentation. However, the improvement in drivable area segmentation and lane detection under the full setting is not obvious, since there are less unlabeled data on these tasks.

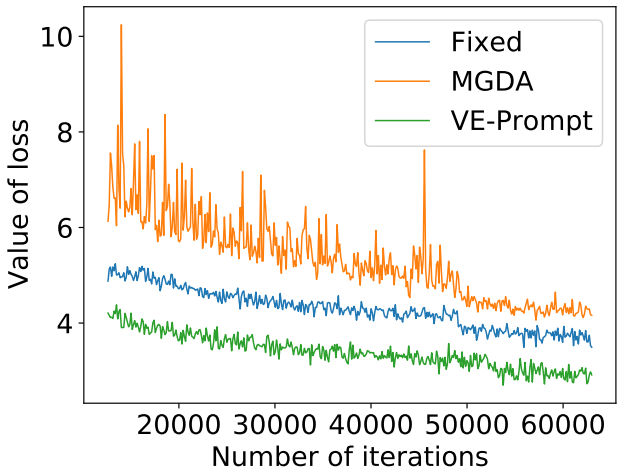

Figure 1. We provide the loss changes of all tasks during training under the disjoint-balance setting. From the curve, we find out that our VE-Prompt can achieve faster and better convergence.

2.3. Task Balancing

We choose pseudo labeling as the baseline since task-balancing methods are more suitable in settings with complete labels. Fixed denotes fixed loss weights for all tasks during training. As shown in Table 2 in the main paper, Uncertainty [6] performs better than Fixed [3] under the full settings overall, while the performance of GradNorm [1] degrades significantly. Interestingly, Fixed performs slightly better than Uncertainty under the disjoint-balance setting, which indicates that Uncertainty is not suitable for all data split settings. GradNorm and MGDA [11] perform poorly overall, showing that these task-balancing methods are not suitable for autonomous driving. Especially, GradNorm uses the last shared layer of weights to compute gradient norm in its paper, thus we adopt the last layer of P_5 in the neck. We also choose the whole shared encoder to implement GradNorm, which is denoted as GradNorm*, improving the original one by a large margin under the disjoint-balance setting (+10.8 in Avg.). This indicates that the selection of network weights for computing GradNorm is important. Interestingly, MGDA consistently achieves the best result on lane detection, indicating that it suffers from the heavy negative transfer. Since it takes a long time to train MGDA, we did not implement it under the full setting for the time limit.

In summary, most existing multi-task learning methods suffer from poor performances under the real-world scenarios of autonomous driving since they are not designed to handle unified perception in self-driving. Therefore, it is extremely important and urgent to develop applicable multi-task methods for autonomous driving.

Table 1. Compare with LV-Adapter under the disjoint-balance setting.

Method	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)	IoU (LD)	GFLOPs	Params
LV-Adapter [8]	24.6	47.4	21.9	**61.8**	80.6	-	415	200M
VE-Prompt (Ours)	**26.8**	**51.2**	**23.8**	58.3	**86.8**	**22.1**	401	60M

Table 2. Comparison of fixed and trainable task-specific prompts under the disjoint-balance setting.

Fixed	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)	IoU (LD)
✓	33.3	55.3	32.5	61.1	87.2	22.1
✗	**33.9**	**56.6**	**33.7**	**61.2**	**87.4**	**22.2**

3. Compare with LV-Adapter

We conduct experiments to compare with recent LV-Adapter [8], which tackles three tasks, as in Table 1. The class number of object detection is 10 in LV-Adapter, and the backbone is Res50 [4]. Here we use the same setting as in LV-Adapter, and present results under the disjoint-balance setting in Table 1. Note that data splits of object detection, semantic segmentation, and drivable area segmentation are the same as in LV-Adapter. Results show that our proposed VE-Prompt performs better than LV-Adapter on object detection and semantic segmentation by a large margin (+2.2 in mAP and +6.2 in mIoU (DA)). Meanwhile, our method gets competitive results in lane detection compared with the Swin-Tiny [10] backbone as in Table 3 in the main paper. LV-Adapter adopts MaskFormer [2], which is a stronger baseline, to generate pseudo labels for semantic segmentation, while VE-Prompt chooses Semantic FPN [7] as the teacher model. Therefore, the improvement of semantic segmentation for LV-Adapter may come from high-quality pseudo labels. The number of parameters in the proposed VE-Prompt is much less than that of LV-Adapter as in Table 1. We also report GFLOPs on the same V100 NVIDIA GPU for a fair comparison. Results show that our method is more efficient and effective overall.

4. More Ablation Studies

4.1. Influence of Fixed Prompts

The task-specific prompts are not fixed during training in VE-Prompt. We also conduct experiments to verify the effectiveness of trainable task-specific prompts as in Table 2. Results show that the model with trainable task-specific prompts performs better on all four tasks.

4.2. Number of Exemplars

Here we compare different configurations of the number of visual exemplars. The number of visual exemplars for different tasks is n_1, n_2, n_3, and n_4. We keep them equal for simplification. As shown in Figure 2, the model performs better when $n_1 = n_2 = n_3 = n_5 = 5$, thus we set the number of visual exemplars as 5 in our final model.

4.3. Loss Analysis

We also analyze the loss changes of VE-Prompt and the baseline under the disjoint-balance setting as in Figure 1. From the loss curves, we conclude that our VE-Prompt achieves consistent faster and better convergence during training. Note that loss weights for all tasks in Fixed and VE-Prompt here are set as 1 for a fair comparison.

4.4. Comparisons with Alternative Options

We present the results of VE-Prompt with some alternative options as in Table 3. Results show that VE-Prompt with Uncertainty can improve Uncertainty on all tasks, and VE-Prompt with Fixed performs better than VE-Prompt with Uncertainty.

Table 3. Results of VE-Prompt with alternative options under the disjoint-balance setting.

Model	mAP	mIoU (SS)	mIoU (DA)	IoU (LD)	$\Delta_{MTL}(\%)$
Uncertainty	31.2	59.9	87.0	22.2	+1.66
VE-Prompt (Uncertainty)	**32.9**	**60.6**	**87.9**	**22.5**	**+4.04**
Fixed	31.3	60.2	87.0	**22.2**	+1.87
VE-Prompt (Fixed)	**33.9**	**61.2**	**87.4**	**22.2**	**+4.72**

5. Experiments on NuImages Dataset

We also conduct experiments on nuImages dataset[1], which covers two tasks, object detection and semantic segmentation. Results are shown in Table 4, indicating that VE-Prompt performs much better than baselines.

References

[1] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for

[1] https://www.nuscenes.org/nuimages

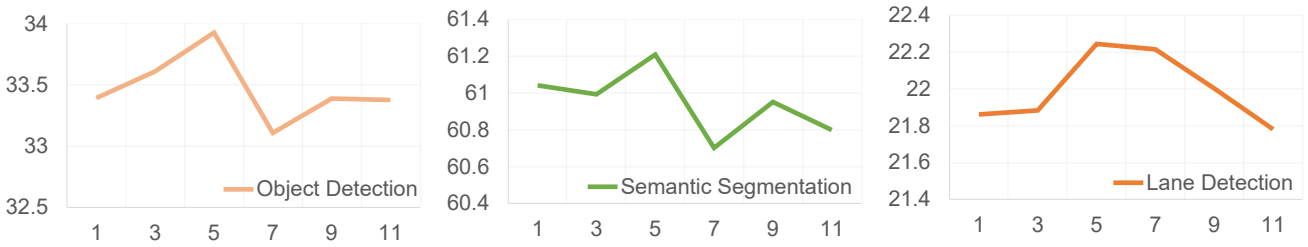

Figure 2. Ablation study of different numbers of visual exemplars under the disjoint-balance setting. The x-axis represents the number of visual exemplars, and the y-axis indicates mAP or mIoU.

Table 4. Comparisons with multi-task baselines on nuImages.

Model	mAP	AP50	AP75	mIoU	Avg.
Sparse R-CNN based	50.4	76.8	54.5	53.8	52.1
DINO based	55.5	81.6	60.6	56.7	56.1
VE-Prompt (Ours)	**55.8**	**81.9**	**60.7**	**59.1**	**57.5**

adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR, 2018. 1

[2] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[3] Golnaz Ghiasi, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8856–8865, 2021. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1013–1021, 2019. 1

[6] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. 1

[7] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, 2019. 2

[8] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. *arXiv preprint arXiv:2209.08953*, 2022. 2

[9] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021. 1

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[11] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 1

[12] Dong Wu, Manwen Liao, Weitian Zhang, and Xinggang Wang. Yolop: You only look once for panoptic driving perception. *arXiv preprint arXiv:2108.11250*, 2021. 1

[13] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 1

[14] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1