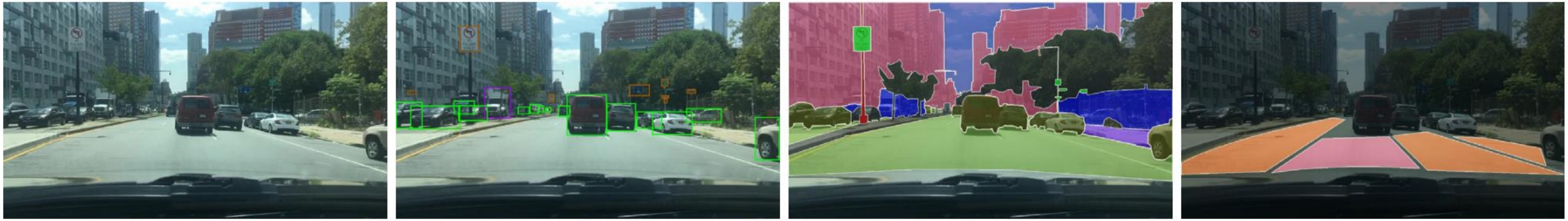
Effective Adaptation in Multi-Task Co-Training for Unified Autonomous Driving

Xiwen Liang[1*], Yangxin Wu[1*], Jianhua Han[2], Hang Xu[2], Chunjing Xu[2], Xiaodan Liang[1†]
[1]Shenzhen Campus of Sun Yat-Sen University, [2]Huawei Noah's Ark Lab
{liangxw29@mail2, wuyx29@mail2, liangxd9@mail}.sysu.edu.cn,
{hanjianhua4, xu.hang, xuchunjing}@huawei.com

Presented by Weiqi

Motivation

- No universal pre-training & pre-training is resource-intensive
- A bridge between pre-training and finetuning



(a) Input (b) Object detection (c) Semantic segmentation (d) Drivable area segmentation

Figure 1: Our multi-task model takes (a) an RGB image as input and tackles (b) traffic object detection, (c) semantic segmentation, and (d) drivable area segmentation simultaneously.

Setting

- Holistic understanding of multiple downstream tasks
  - Tasks: semantic & drivable area segmentation, traffic object detection
  - Extracting features with better **transferability**
  - Pretrain-finetune V.S. pretrain-adapt-finetune
    - Poor performance due to **distinction of training objectives & architecture design**
    - **Prompt-based adaption:** learnable task-specific prompting and L2V alignment

- Novelty
  - **Adapt stage** & unchanged backbone (few training cost in FPN)
  - LV-Adapter benefits downstream tasks

Setting

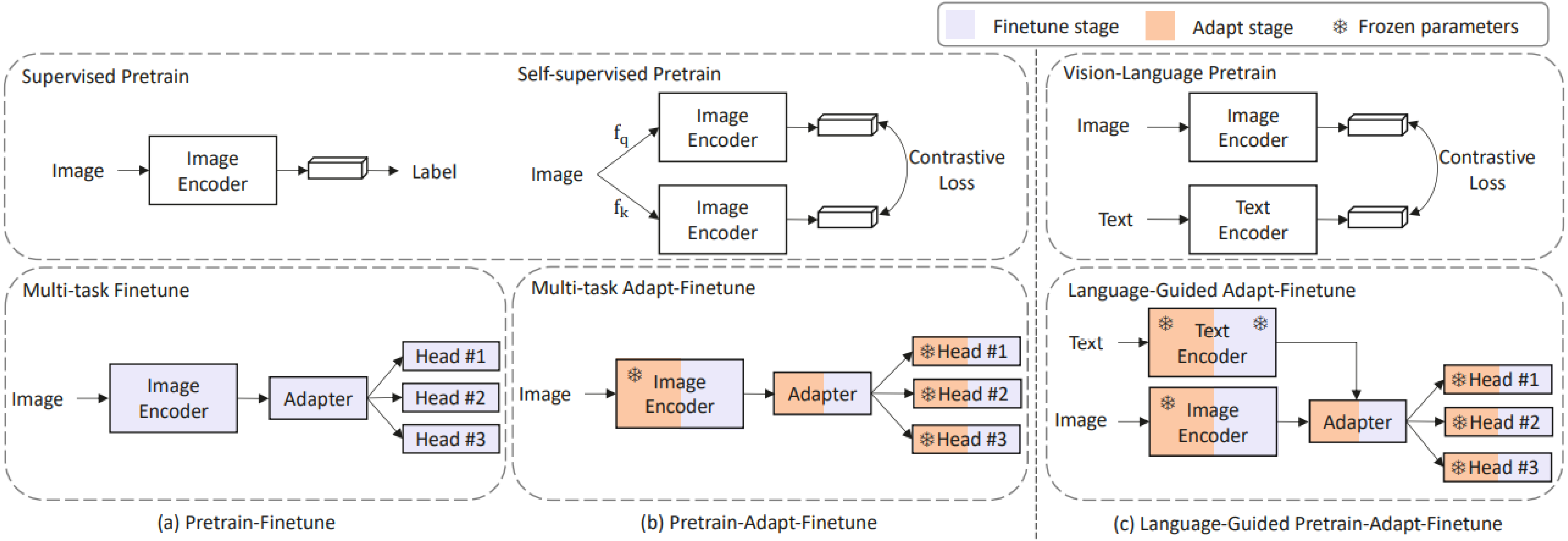


Figure 2: Comparisons of the conventional *pretrain-finetune* paradigm and our proposed *pretrain-adapt-finetune* paradigm. The language-guided *pretrain-adapt-finetune* paradigm further incorporates language priors into multiple downstream tasks.

Pipeline

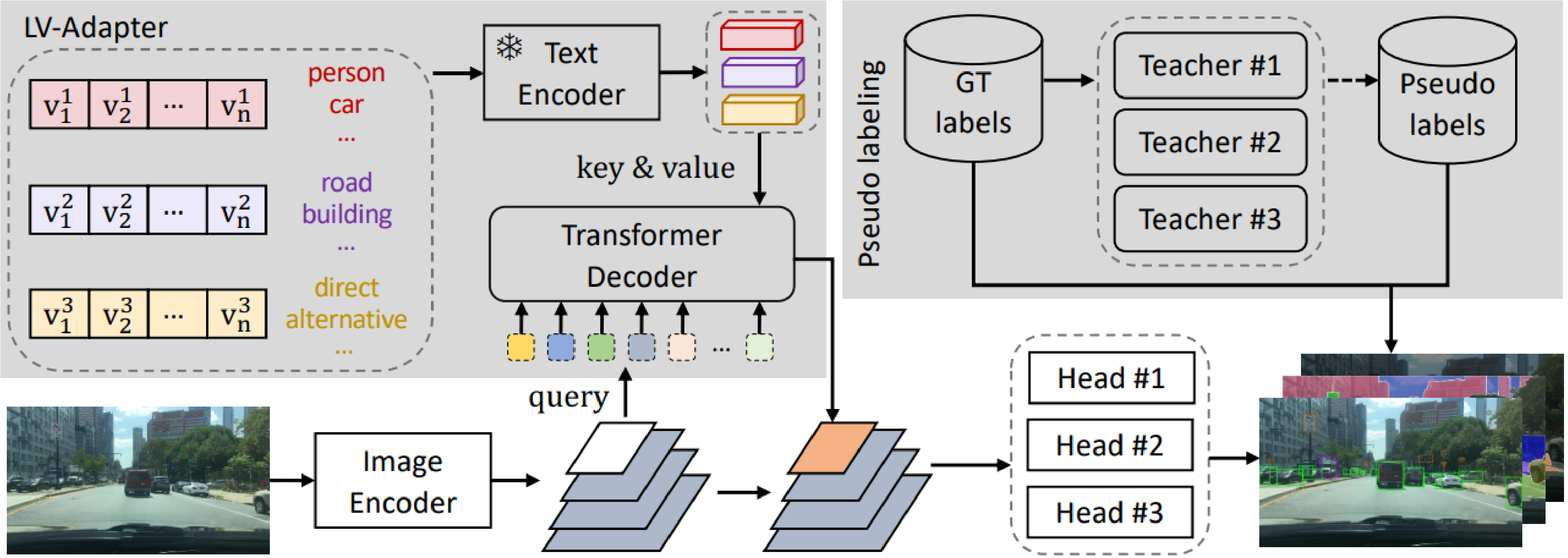


Figure 3: An overview of our proposed model. We first train three specialized teacher on labeled data to generate pseudo labels for each task. The multi-task model is then trained on both ground-truth and pseudo labels under the language-guided *pretrain-adapt-finetune* paradigm.

Details

- Each task share the same backbone with task-specific head
  - Backbone&neck:
    - ResNet, FPN
  - Seg Head:
    - MaskFormer
  - Det Head
    - Sparse R-CNN

Adapt stage

Teacher model to generate pseudo labels:

- Teacher: trained with few labeled for each task
- Pseudo label: prediction of unlabeled data by teacher

Student model trained with both gTruth and pseudo label：

- Language-guided pretrain-adapt-finetune paradigm:

$$\mathcal{L}_{total} = \alpha_{det}\mathcal{L}_{det} + \alpha_{sem}\mathcal{L}_{sem} + \alpha_{driv}\mathcal{L}_{driv}$$

L2V: language guided

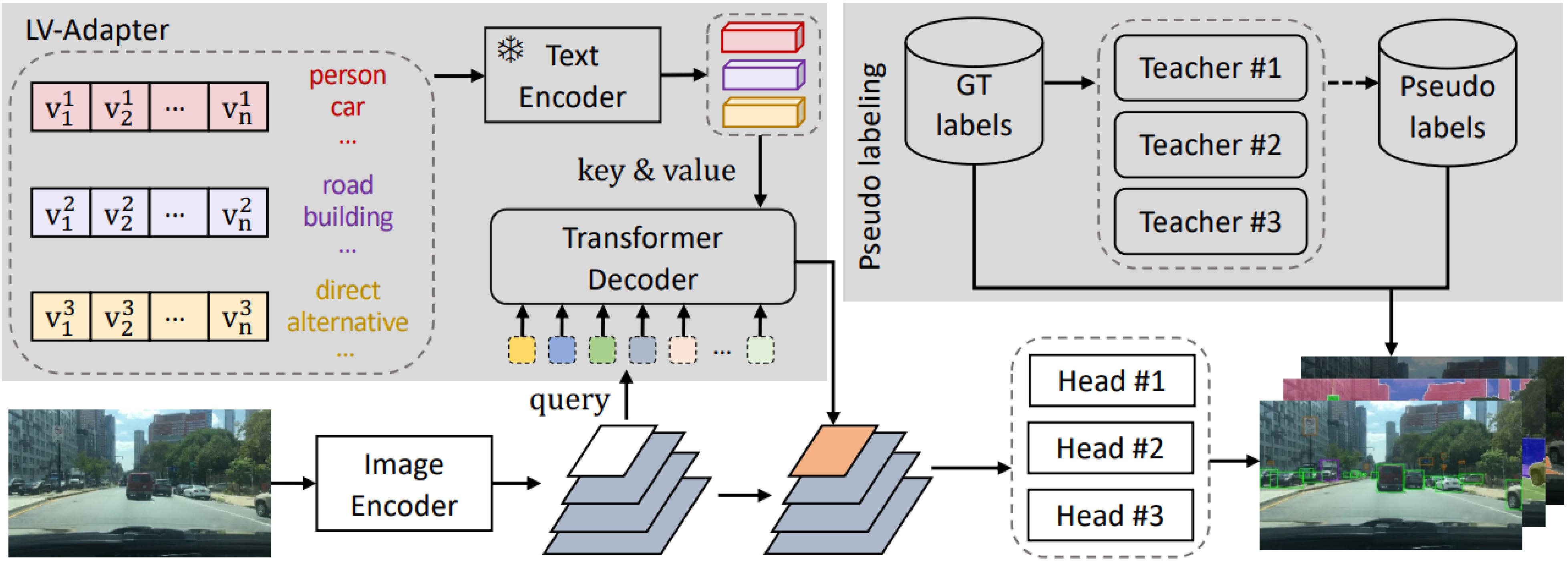


Figure 3: An overview of our proposed model. We first train three specialized teacher on labeled data to generate pseudo labels for each task. The multi-task model is then trained on both ground-truth and pseudo labels under the language-guided *pretrain-adapt-finetune* paradigm.

L2V: language guided

- LV-Adapter
  - Task-specific prompts: learnable template(from CoOp) for each task

$$\hat{\mathcal{T}}_{e,i} = \text{L2\_NORM}(\text{TE}([\mathbf{v}^t, \mathbf{n}_i])$$

    - n_i: embedding of the class name
    - vt: task-specific learnable contexts t-task

  - Language-aware context(cross-attention)

$$\mathcal{A}_{L\to V}(\hat{\mathcal{T}}_e, \mathbf{z}_5) = \text{TransDecoder}(q = \mathbf{z}_5, k = \hat{\mathcal{T}}_e, v = \hat{\mathcal{T}}_e),$$

    - Z_5: last feature map of P5 (strides=32 pyramidal feature from FPN)
    - Output $\underline{Z_5}$ to update Z_5, thus FPN is linguistic-aware, and then connected to task-specific heads.

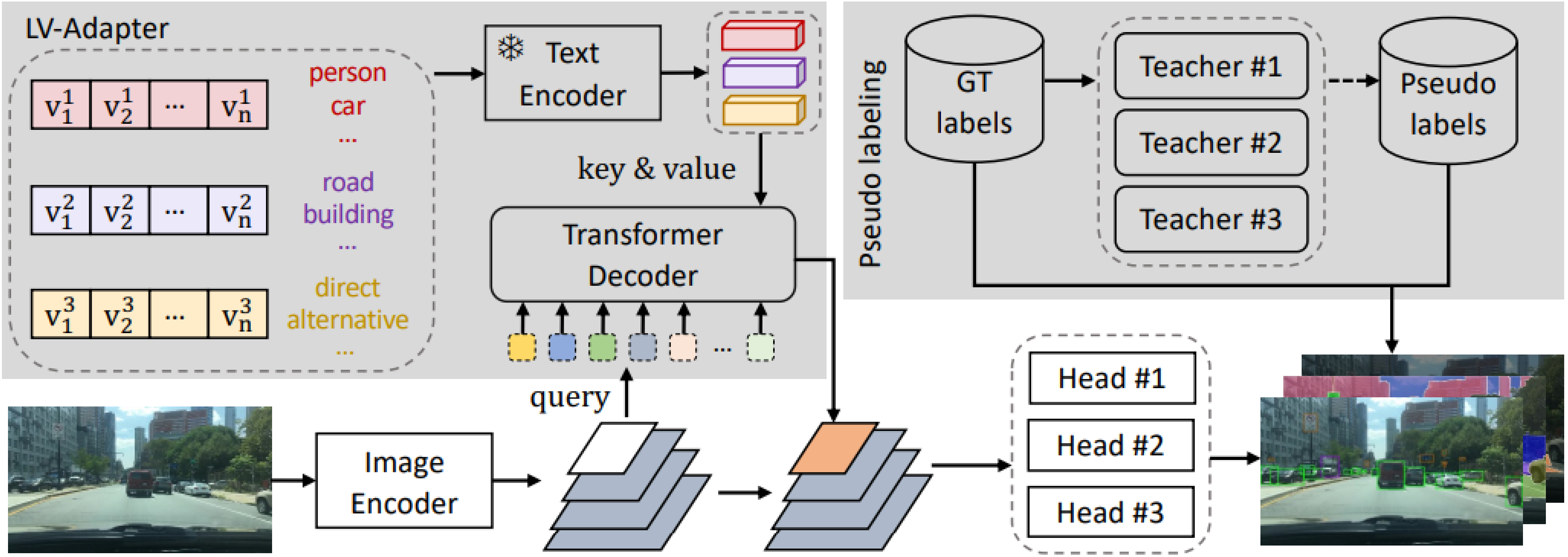


Figure 3: An overview of our proposed model. We first train three specialized teacher on labeled data to generate pseudo labels for each task. The multi-task model is then trained on both ground-truth and pseudo labels under the language-guided *pretrain-adapt-finetune* paradigm.

Experiment

- BDD100K
  - 100k videos of 40s, annotations at a flip of second 10
  - 67K training images: object detection and drivable area segmentation (no ss)
  - 4k training images: semantic segmentation (no od, das)
  - 3K training images: object detection, drivable area segmentation, semantic segmentation

- Disjoint-normal
  - drivable area segmentation (20k), object detection (10k), semantic segmentation (7k).

Experiment

- Comparation of multi-task training methods

Table 1: Comparisons of popular multi-task training methods and our proposed LV-Adapter under the Disjoint-normal setting.

Method	mIoU (SS)	mIoU (DA)	mAP	AP50	AP75
Zeroing loss [46]	59.4	83.3	23.0	46.0	19.8
Uniform sampler [24]	59.9	83.2	24.3	47.0	21.8
Weighted sampler [24]	59.8	83.2	24.2	46.9	21.4
Round-robin [24]	60.7	83.1	24.2	47.0	21.5
Self-training [13]	60.3	83.1	24.9	48.1	22.2
LV-Adapter (Ours)	**62.2**	**83.7**	**26.4**	**50.5**	**23.7**

$$\mathcal{L}_{total} = \alpha_{det}\mathcal{L}_{det} + \alpha_{sem}\mathcal{L}_{sem} + \alpha_{driv}\mathcal{L}_{driv}$$

Table 2: Comparisons of different paradigms under the Disjoint-normal setting with ResNet-50 backbone. Orange color indicates the results of our proposed *pretrain-adapt-finetune* paradigm, while others are results of conventional *pretrain-finetune* paradigm.

		Semantic Seg.		Drivable Seg.		Object Detection		
Type	Model	mIoU	pACC	mIoU	pACC	mAP	AP50	AP75
Classification-oriented	MoCo-v1 [17]	17.8	48.6	70.8	92.0	25.8	49.5	23.1
		59.2	93.2	83.6	96.9	25.9	50.0	23.0
	MoCo-v2 [4]	10.3	19.8	73.4	93.5	26.0	50.1	23.2
		61.2	93.4	83.8	96.9	26.1	50.4	23.4
	SimCLR [3]	60.3	93.3	83.5	96.8	25.4	48.9	22.5
		60.1	93.2	83.5	96.8	25.2	48.9	22.3
	SwAV [2]	45.9	71.1	82.0	96.5	25.6	49.1	23.0
		61.1	93.3	83.1	96.7	25.6	49.3	23.1
	BYOL [15]	59.2	90.2	75.6	93.9	25.9	49.8	23.4
		61.7	93.4	83.5	96.8	25.7	49.4	23.1
Detection-oriented	DetCo [47]	38.1	58.5	83.2	96.7	25.9	49.7	22.9
		61.0	93.4	83.7	96.9	26.2	50.3	23.3
Segmentation-oriented	DenseCL [44]	20.0	40.0	73.7	93.7	26.1	50.3	23.5
		60.7	93.3	83.9	96.9	26.3	50.3	23.7
Vision-language	CLIP [32]	54.5	91.1	74.1	93.1	26.5	50.7	23.8
		61.0	93.2	83.4	96.8	26.3	50.5	23.5

where $\mathcal{L}_{det}, \mathcal{L}_{sem}, \mathcal{L}_{driv}$ are losses for object detection, semantic segmentation, and drivable area segmentation, respectively. [1] We include more details in Section 5.1.

Single task v.s. mutil-task training

Table 3: Results of single-task baselines and multi-task models with ResNet-50 backbone. SS and DA means semantic segmentation and drivable area segmentation. - indicates inapplicable.

Setting	Method	mIoU (SS)	mIoU (DA)	mAP	AP50	AP75
Full	MaskFormer [5]	57.1	-	-	-	-
	MaskFormer [5]	-	83.9	-	-	-
	Sparse R-CNN [38]	-	-	29.4	55.8	26.4
	Self-training [13]	61.8	84.4	30.1	56.6	27.6
	LV-Adapter (Ours)	**63.1**	**84.9**	**31.1**	**58.2**	**28.4**
Disjoint-balance	MaskFormer [5]	57.1	-	-	-	-
	MaskFormer [5]	-	78.1	-	-	-
	Sparse R-CNN [38]	-	-	18.6	37.8	15.6
	Self-training [13]	59.4	80.3	22.4	44.1	19.6
	LV-Adapter (Ours)	**61.8**	**80.6**	**24.6**	**47.4**	**21.9**
Disjoint-normal	MaskFormer [5]	57.1	-	-	-	-
	MaskFormer [5]	-	82.0	-	-	-
	Sparse R-CNN [38]	-	-	20.9	41.9	17.8
	Self-training [13]	60.3	83.1	24.9	48.1	22.2
	LV-Adapter (Ours)	**62.2**	**83.7**	**26.4**	**50.5**	**23.7**

Ablation

Table 4: Ablation study of the components of our proposed LV-Adapter.

#	Prompt	V2L	L2V	mIoU (SS)	mIoU (DA)
1	✗	✗	✗	54.5	74.1
2	✗	✗	✗	61.5	83.5
3	✓	✗	✗	61.8	83.5
4	✓	✓	✗	61.3	83.6
5	✓	✗	✓	**62.2**	**83.7**

Training efficiency

Table 5: Comparison of different configurations of *adapt* and *finetune* epochs.

Adapt	Finetune	mIoU (SS)	mIoU (DA)	mAP
12	24	59.5	82.9	25.8
6	30	61.0	83.4	26.3
1	35	61.5	83.5	26.4
6	35	61.3	83.5	26.5
12	35	61.9	83.5	26.6

Prompts

Table 6: Comparison with handcrafted prompts under the disjoint-normal setting.

Method	mAP	AP50	AP75	mIoU (SS)	mIoU (DA)
Prompt engineering	26.1	50.1	23.4	61.4	83.3
Prompt ensembling	26.3	50.4	23.5	61.7	83.3
Learned prompts	**26.4**	**50.5**	**23.7**	**62.2**	**83.7**

All convolutional based exp.

Table 8: Comparisons of different training schemes for self-supervised models with convolutional heads.

Method	Scheme	mIou (SS)	mIoU (DA)	mAP
SimCLR	pretrain-finetune	**58.3**	83.2	24.0
	pretrain-adapt-finetune	**58.3**	**83.3**	**24.6**
DetCo	pretrain-finetune	59.6	83.3	24.1
	pretrain-adapt-finetune	**59.7**	**83.4**	**24.8**

Single-task setting

Table 10: Comparisons on single-task setting.

Model	mIoU (SS)	mIoU (DA)	mAP
MaskFormer	57.1	-	-
+ LV-Adaper	**61.3**$^{+4.2\%}$	-	-
MaskFormer	-	78.1	-
+ LV-Adaper	-	**81.6**$^{+3.5\%}$	-
Sparse R-CNN	-	-	18.6
+ LV-Adaper	-	-	**22.4**$^{+3.8\%}$

Experiment on NuImages

Table 11: Comparisons of single-task and multi-task models on NuImages dataset.

Method	mAP	AP50	AP75	mIoU
Sparse R-CNN	46.6	72.8	49.4	-
MaskFormer	-	-	-	55.8
Multi-task	47.1	73.5	49.9	53.3
LV-Adapter	**50.3**	**76.8**	**54.3**	**56.0**

Summary

- Pretrain-adapt-finetune paradigm
  - Reduce the gap between pre-training and fine-tuning without increasing training cost
- LV-Adapter
  - Incorporate linguistic knowledge into visual features which benefit the downstream tasks
- A more diversified task set with more annotations will further benefit the learning of multi-headed architecture