
Object Concepts Emerge from Motion

Supplementary Material

Anonymous Author(s)

Affiliation

Address

email

A Pseudo-codes for Pixel Cluster

For all optical flow data generated by VideoFlow, we perform a simple Breadth-First Search(BFS) to segment moving objects. Alg. 1 provides a pseudocode description of our algorithm. The algorithm takes the optical flow, the forward-backward consistency check result, and two thresholds θ_f and θ_s as input. θ_f is used to determine when the optical flow of two adjacent pixels, being sufficiently close, is considered to belong to the same object. θ_s controls the minimum number of pixels that an object should have.

B Data Augmentation Details

All input images are first randomly resized to a resolution between 512×288 and 1024×576 . They are then randomly cropped to 224×224 . During cropping, up to 10 attempts are made to ensure that the cropped region contains at least two distinct labels. Afterward, each image has a 50% chance of being horizontally flipped. Additionally, gamma, brightness, and color augmentations are applied with a 50% probability, each sampled within the range of (0.9, 1.1).

C Monocular Depth Estimation on KITTI Official Leaderboard

Tab. 1 shows the results on the official KITTI online benchmark. Our method outperforms previous methods in the primary metric (SILog) and also achieves competitive performance across the other evaluation metrics.

Table 1: Quantitative results on the official split of KITTI dataset. All metrics reported here are from the KITTI online leaderboard.

Method	Backbone	Pretrain	SILog ↓	Abs Rel ↓	Sq Rel ↓	iRMSE ↓
NeW CRFs [9]	Swin-Large	ImageNet sup.	10.39	8.37	1.83	11.03
VA-DepthNet [2]	Swin-Large	ImageNet sup.	9.63	7.96	1.66	10.44
IEBins [5]	Swin-v2-Large	MIM [7]	9.84	7.82	1.60	10.68
NDDepth [4]	Swin-v2-Large	MIM [7]	9.62	7.75	1.59	10.62
DCDepth [6]	Swin-Large	Semantic-SAM [1]	<u>9.60</u>	7.83	1.54	10.12
DCDepth [6]	Swin-Large	Ours	9.54	<u>7.76</u>	<u>1.55</u>	<u>10.37</u>

D 3D Object Detection on nuScenes val Set

To compare with more methods based on ViT architectures whose cost are not affordable for large input resolution, we also test various methods with image size of 704×256 . Tab. 2 presents the results of BEVFormerV2 [8] on the nuScenes val set utilizing various pretrained image backbones. All experiments use two input frames providing temporal information. Our method consistently outperforms the baselines pretrained on ImageNet-22K across all evaluated backbone architectures. Compared to DINOv2, our Swin-based models achieve competitive or superior performance with

Algorithm 1 Pixel Cluster

Input: flow(optical flow), valid(consistency check), θ_f, θ_s

```
1: Initialization:  $n \leftarrow 0, v[i][j] \leftarrow false, S \leftarrow \emptyset$ 
2: for  $x \leftarrow 1$  to  $H$  do
3:   for  $y \leftarrow 1$  to  $W$  do
4:     if  $v[x][y] = true$  or  $valid[x][y] = false$  then
5:       continue
6:     end if
7:      $Q \leftarrow$  empty queue,  $C \leftarrow \emptyset$ 
8:     Enqueue( $Q, (x, y)$ )
9:     while  $Q \neq \emptyset$  do
10:       $(x, y) \leftarrow$  Dequeue( $Q$ )
11:       $C \leftarrow C \cup \{(x, y)\}$ 
12:      for  $(i, j)$  in  $(x, y)$ 's 4 neighbors do
13:        if  $\|flow[i][j], flow[x][y]\|_2 \leq \theta_f$  and  $v[i][j] = false$  and  $valid[x][y] = true$  then
14:           $v[i][j] = true$ 
15:          Enqueue( $Q, (i, j)$ )
16:        end if
17:      end for
18:    end while
19:    if  $|C| \geq \theta_s$  then
20:       $S \leftarrow S \cup \{C\}$ 
21:    end if
22:  end for
23: end for
Output:  $S$ 
```

Table 2: Quantitative evaluation of BEVFormerV2 [8] on nuScenes val set using different pretraining methods.

Method	Backbone	NDS \uparrow	mAP \uparrow	mATE \downarrow	mASE \downarrow	mAOE \downarrow	mAVE \downarrow	mAAE \downarrow
DINOv2	ViT-S	0.4624	0.3488	0.7165	0.2845	0.4997	0.4270	0.1884
DINOv2	ViT-B	0.4908	0.3836	0.6974	0.2821	0.4181	0.4240	0.1884
DINOv2	ViT-L	0.5191	0.4205	0.6504	0.2735	0.3645	0.4379	0.1851
ImageNet-22K	Swin-T	0.4742	0.3634	0.7090	0.2840	0.4836	0.4047	0.1940
Ours	Swin-T	0.4824	0.3708	0.7061	0.2799	0.4445	0.4054	0.1945
ImageNet-22K	Swin-S	0.4878	0.3800	0.7065	0.2835	0.4120	0.4295	0.1901
Ours	Swin-S	0.5087	0.4023	0.6888	0.2785	0.4045	0.3667	0.1861
ImageNet-22K	Swin-B	0.5042	0.4071	0.6856	0.2780	0.4060	0.4281	0.1952
Ours	Swin-B	0.5169	0.4136	0.6601	0.2803	0.3798	0.3973	0.1810
ImageNet-22K	Swin-L	0.5048	0.4009	0.6801	0.2790	0.4069	0.4067	0.1838
Ours	Swin-L	0.5203	0.4179	0.6610	0.2812	0.3684	0.4013	0.1751

significantly fewer parameters and computational cost. For instance, our model pretrained with the Swin-L backbone attains an NDS of **0.5203** and an mAP of **0.4179**. These results are comparable to those achieved by DINOv2 with the ViT-L backbone.

E More Results on 3D Occupancy Perception

Tab. 3 presents the results of SparseOcc [3] on the nuScenes val set with more pretrained image backbones. Notably, our model pretrained with the Swin-L backbone achieves a RayIoU of **38.7**, which is competitive with the **39.0** RayIoU obtained by DINOv2 using the larger ViT-L backbone. Furthermore, our Swin-T model achieves a RayIoU of 37.0, outperforming the DINOv2 ViT-S model (35.9) and performing competitively with the DINOv2 ViT-B model (37.1).

F More Qualitative Results

Fig. 1 shows additional qualitative results of the pseudo-label generation and the visualizations of the output features. As illustrated in the pseudo-label visualizations, the proposed algorithm successfully

Table 3: Quantitative evaluation of SparseOcc [3] on nuScenes val set using different pretraining methods.

Method	BackBone	RayIoU	RayIoU _{1m, 2m, 4m}		
DINOv2	ViT-S	35.9	29.5	36.8	41.4
DINOv2	ViT-B	<u>37.1</u>	<u>31.0</u>	<u>37.9</u>	<u>42.4</u>
DINOv2	ViT-L	39.0	32.8	39.9	44.3
Ours	Swin-T	37.0	31.1	37.8	42.2
Ours	Swin-S	38.1	32.0	<u>39.1</u>	43.4
Ours	Swin-B	<u>38.3</u>	<u>32.1</u>	<u>39.1</u>	<u>43.7</u>
Ours	Swin-L	38.7	32.6	39.5	43.8

segments objects exhibiting significant movement, as well as foreground instances exhibiting motion patterns distinct from the background. The feature visualizations shows that the model distinguishes many objects not annotated in the pseudo-labels. This suggests our model goes beyond mimicking pseudo-labels, but learning a more general, object-centric representation.

References

- [1] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *ECCV*, 2024.
- [2] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. VA-DepthNet: A variational approach to single image depth prediction. In *ICLR*, 2023.
- [3] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3D occupancy prediction. In *ECCV*, 2024.
- [4] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. NDDepth: Normal-distance assisted monocular depth estimation. In *ICCV*, 2023.
- [5] Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. IEBins: Iterative elastic bins for monocular depth estimation. In *NeurIPS*, 2023.
- [6] Kun Wang, Zhiqiang Yan, Junkai Fan, Wanlu Zhu, Xiang Li, Jun Li, and Jian Yang. DCDepth: Progressive monocular depth estimation in discrete cosine domain. In *NeurIPS*, 2024.
- [7] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *CVPR*, 2023.
- [8] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. BEVFormer v2: Adapting modern image backbones to bird’s-eye-view recognition via perspective supervision. In *CVPR*, 2023.
- [9] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected CRFs for monocular depth estimation. In *CVPR*, 2022.

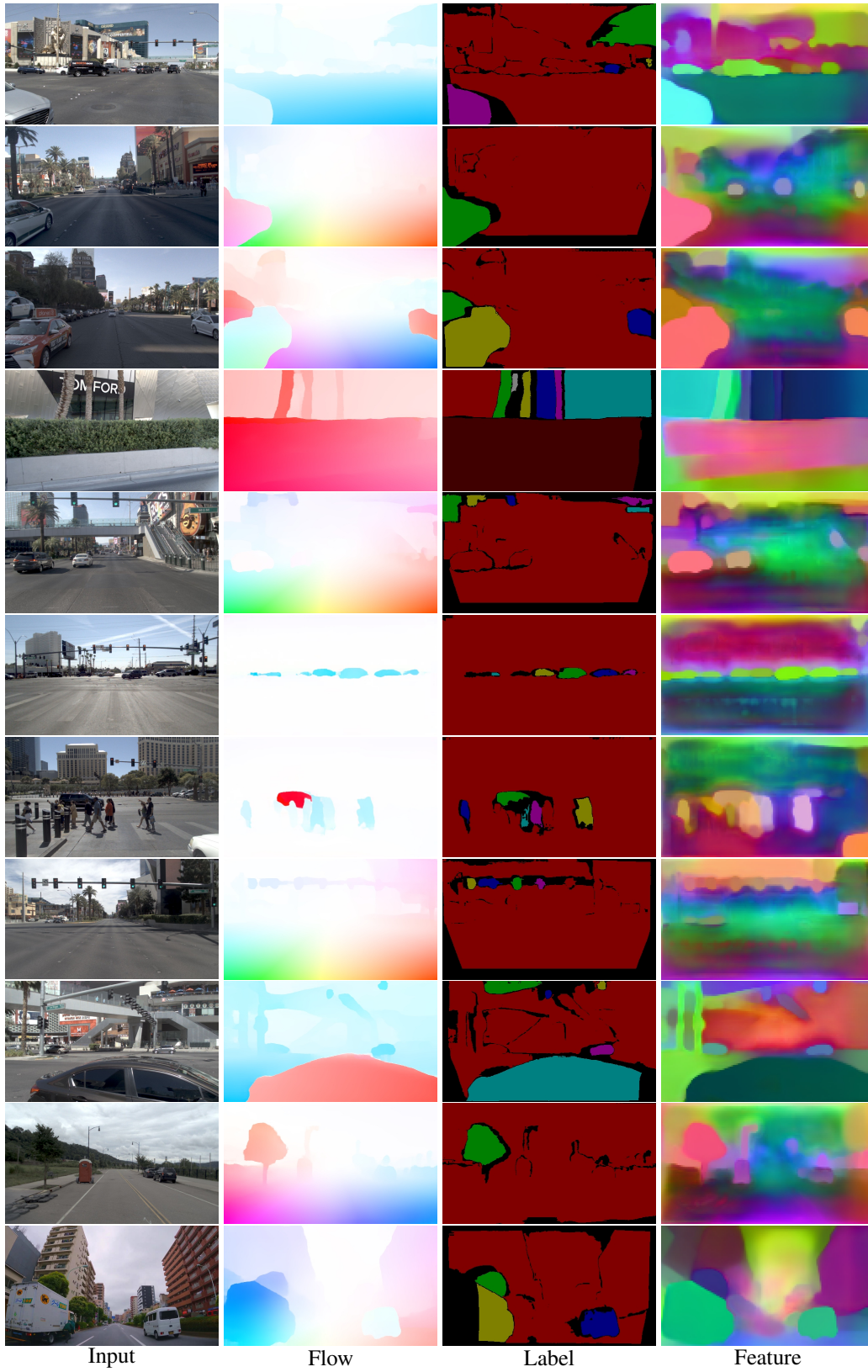


Figure 1: Examples of the pseudo-label generation results and the output features.