

Textual Query-Driven Mask Transformer for Domain Generalized Segmentation

Supplementary Material

Appendix

In Appendix, we provide more details and additional experimental results of our proposed tqdm. The sections are organized as follows:

- A. Text Activation in Diverse Domains
- B. Details of Motivating Experiment
- C. Experiment on SYNTHIA Dataset
- D. Details of Region Proposal Experiment
- E. More Qualitative Results
- F. Qualitative Results on Unseen Game Videos
- G. Comparison with Open-Vocabulary Segmentation

A. Text Activation in Diverse Domains

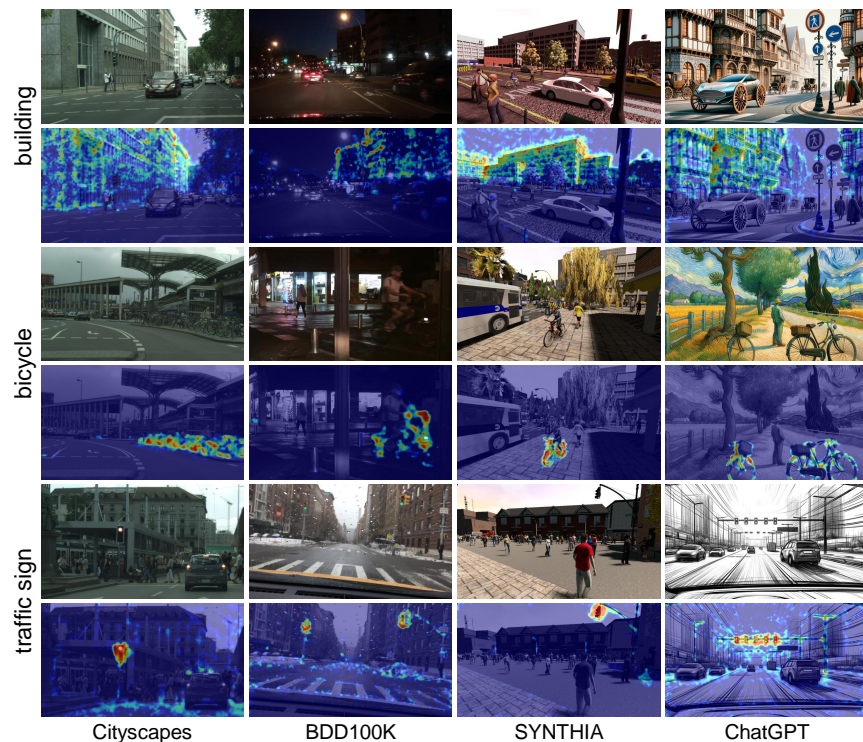


Figure S1: Image-text similarity map on diverse domains. The text embeddings of the targeted classes (*i.e.*, ‘building,’ ‘bicycle,’ and ‘traffic sign’) are consistently well-activated within the corresponding class regions of images across various domains.

We find an interesting property of VLMs: the text embedding of a class name is well-aligned with the visual features of the class region across various domains. Specifically, we visualize the image-text similarity map \mathbf{M} of a pre-trained VLM [8, 19], as shown in Fig. S1. Firstly, we begin by extracting the visual features $\mathbf{x} \in \mathbb{R}^{hw \times C}$ from images across various domains, where h and w are the output resolutions of the image encoder. In parallel, we obtain the text embeddings $\mathbf{t} \in \mathbb{R}^{K \times C}$ for the names of K classes. Following this, we calculate the similarity map $\hat{\mathbf{x}}\hat{\mathbf{t}}^\top$, where $\hat{\mathbf{x}}$ and $\hat{\mathbf{t}}$ are the ℓ_2 normalized versions of \mathbf{x} and \mathbf{t} , respectively, along the C dimension. Lastly, we reshape and resize the resulting similarity map to the original image resolution. During this process, we also normalize the values using min-max scaling. The equation to compute \mathbf{M} is as follows:

$$\mathbf{M} = \text{norm}(\text{resize}(\text{reshape}(\hat{\mathbf{x}}\hat{\mathbf{t}}^\top))). \quad (\text{S1})$$

We adopt the EVA02-CLIP model as a VLM. In Fig. S1, each text embedding (e.g., ‘bicycle’) shows strong activation with the visual features of the corresponding class region across different visual domains. These findings suggest that text embeddings can serve as a reliable basis for domain-invariant pixel grouping.

B. Details of Motivating Experiment

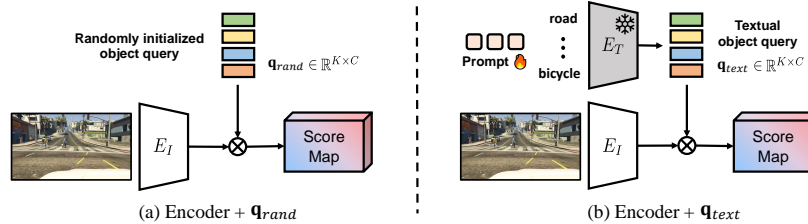


Figure S2: We compare (a) randomly initialized object queries and (b) textual object queries using a simple model architecture, which comprises an image encoder E_I and object queries \mathbf{q} . For the encoder, we use a ViT-base model with CLIP initialization.

In Sec. 3, we demonstrate the superior ability of textual object queries to generalize to unseen domains. As shown in Fig. S2, we compare textual object queries with conventional randomly initialized object queries using a simple model architecture. This architecture comprises an image encoder E_I from a VLM and K object queries $\mathbf{q} \in \mathbb{R}^{K \times C}$. Given the ℓ_2 normalized queries $\hat{\mathbf{q}}$ and visual embeddings $\hat{\mathbf{x}} \in \mathbb{R}^{hw \times C}$ from E_I , the segmentation logits $\mathbf{S} = \hat{\mathbf{x}}\hat{\mathbf{q}}^\top \in \mathbb{R}^{hw \times K}$ are optimized with a per-pixel cross-entropy loss, as described in Eq. 6. In this experiment, we use a CLIP-initialized Vision Transformer-base (ViT-B) backbone with a patch size of 16. The models are trained on GTA5 [12] or SYNTHIA [13], with a crop size of 512×512 , a batch size of 16, and 5k training iterations. The learning rate is set to 1×10^{-4} , and the backbone learning rate is set to 1×10^{-5} .

C. Experiment on SYNTHIA Dataset




Method	S→C	S→B	S→M	Avg.
SAN-SAW [11]	40.87	35.98	37.26	38.04
TLDR [6]	42.60	35.46	37.46	38.51
IBAFormer [14]	50.92	44.66	50.58	48.72
VLTSeg [5] 	56.80	50.50	54.50	53.93
tqdm (ours) 	57.99	52.43	54.87	55.10

Table S1: Comparison of mIoU (%; higher is better) between DGSS methods trained on S and evaluated on C, B, M.  denotes EVA02-CLIP [15] pre-training. The best results are **highlighted** and our method is marked in **blue**.

We conduct an additional experiment in the synthetic-to-real setting (*i.e.*, $S \rightarrow \{C, B, M\}$), and the results are shown in Tab. S1. In this experiment, we train on SYNTHIA [13], and evaluate on Cityscapes [2], BDD100K [18], and Mapillary [10]. Our **tqdm** consistently outperforms other DGSS methods across all benchmarks, demonstrating superior synthetic-to-real generalization capability.

D. Details of Region Proposal Experiment

In this section, we provide a detailed explanation of the experiment on the robustness of object query representation discussed in Sec. 5.3, and present further experimental results in Fig. S3. In Fig. 5, we compare the region proposal results between our **tqdm** and the baseline. Given the per-pixel embeddings $\mathbf{Z} \in \mathbb{R}^{H \times W \times D}$ from the pixel decoder, along with the initial object queries $\mathbf{q}^0 \in \mathbb{R}^{K \times D}$, region proposals $\mathbf{R} \in \{0, 1\}^{H \times W \times K}$ are predicted as follows:

$$\mathbf{R} = \begin{cases} 1, & \text{if } \text{sigmoid}(\mathbf{Z}\mathbf{q}^{0\top}) > \theta \\ 0, & \text{otherwise} \end{cases} \quad (\text{S2})$$

where H and W represent the spatial resolutions, D is the channel dimension, K denotes the number of queries, and θ is defined as a confidence threshold. By incrementally adjusting θ from 0.0 to 1.0, we generate precision-recall curves for each region proposal by class, as depicted in Fig. S3a. Our **tqdm** significantly surpasses the baseline in identifying rare classes (*i.e.*, ‘train,’ ‘motorcycle,’ ‘rider,’ and ‘bicycle’), and achieves marginally better performance across most other classes. Intriguingly, this pattern of enhancement in AP is mirrored in the class-wise IoU results of final predictions, as shown in Fig. S3b. These results suggest that the robustness of query representations for semantic regions plays a crucial role in the generalizability of the final mask predictions that stem from these region proposals.

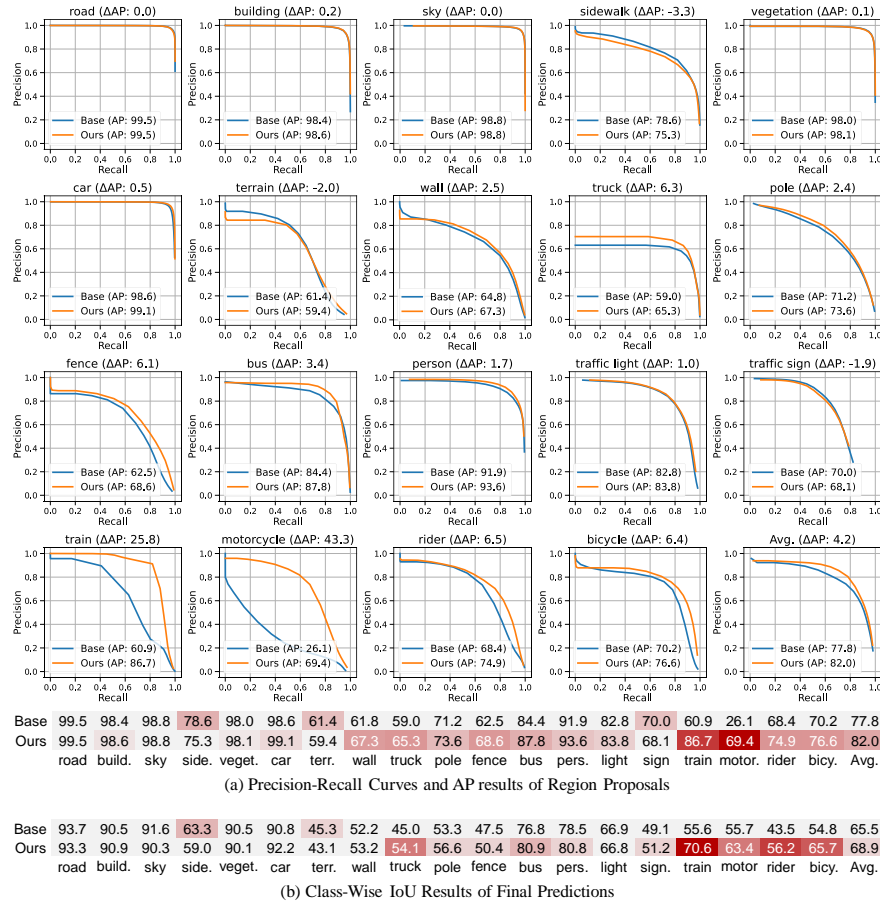


Figure S3: (a) The precision-recall curves and AP results for region proposals. (b) The class-wise IoU results of final predictions. The class-wise trends observed in both tables show a similar pattern.

E. More Qualitative Results

In this section, we provide qualitative comparisons with other DGSS methods [6, 16], which demonstrate superior performance using ResNet [6] and ViT [16] encoders, respectively. All the models are trained on GTA5 [12]. Figs. S4 to S6 display the qualitative results in the synthetic-to-real setting ($G \rightarrow \{C, B, M\}$), respectively. Our **tqdm** yields superior results compared to existing DGSS methods [6, 16] across various domains.

In Fig. S7, we further present qualitative comparisons under extreme domain shifts, showcasing results for hand-drawn images (row 1 and 2), game scene images (row 3 and 4), and images generated by ChatGPT (row 5 and 6). Notably, **tqdm** demonstrates more accurate predictions even under extreme domain shifts, as it is capable of comprehending semantic knowledge.

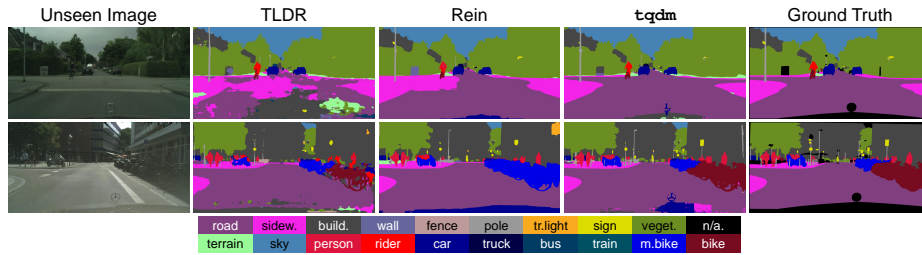


Figure S4: Qualitative results of DGSS methods [6, 16], and our tqdm on G→C.

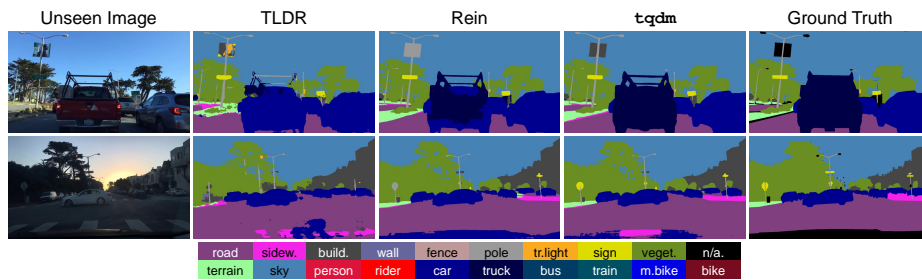


Figure S5: Qualitative results of DGSS methods [6, 16] and our tqdm on G→B.

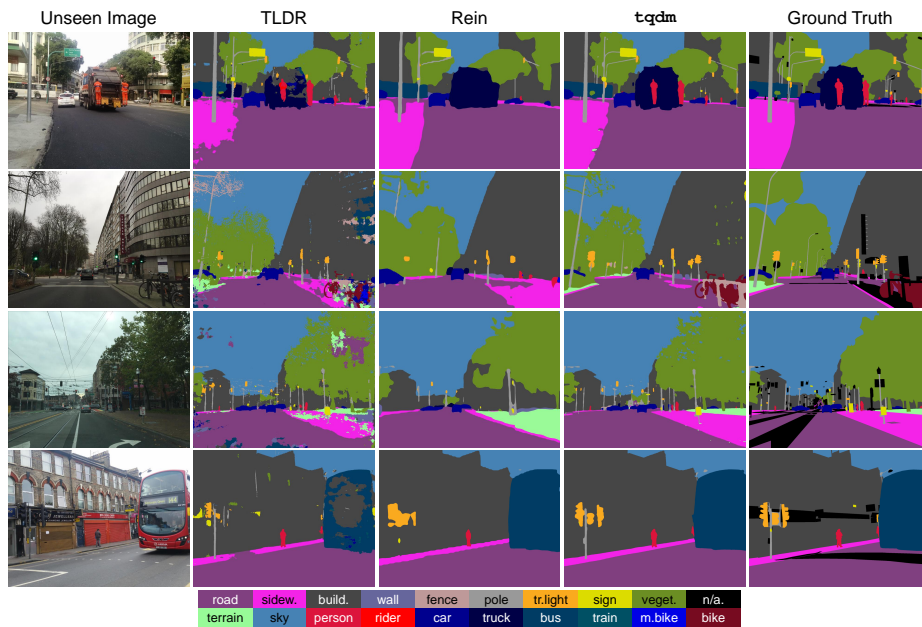


Figure S6: Qualitative results of DGSS methods [6, 16] and our tqdm on G→M.

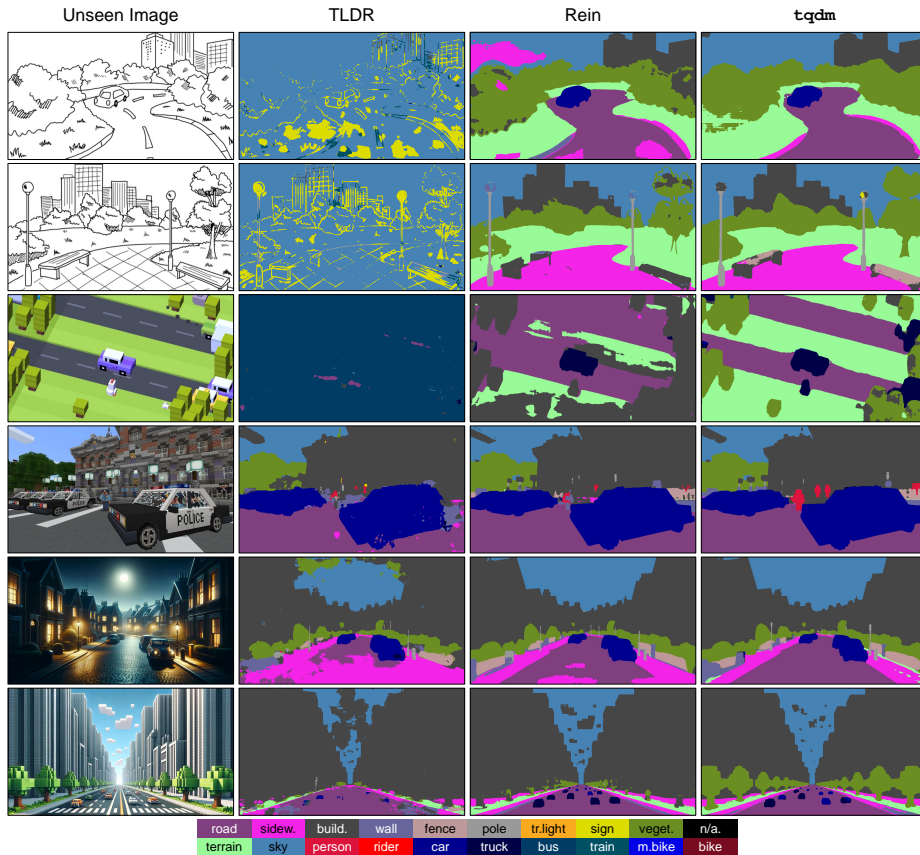


Figure S7: Qualitative results of DGSS methods [6, 16] and our $tqdm$, trained on G and evaluated under extreme domain shifts. We present results for hand-drawn images (row 1 and 2), game scene images (row 3 and 4), and images generated by ChatGPT (row 5 and 6).

F. Qualitative Results on Unseen Game Videos

To ensure more reliable results, we perform qualitative comparisons with other DGSS methods [6, 16] on unseen videos. All the models are trained on GTA5 [12]. Our $tqdm$ consistently outperforms the other methods by delivering accurate predictions in unseen videos. Notably, in both the first and last clips, $tqdm$ effectively identifies trees as the `vegetation` class and clearly distinguishes the `road` and `terrain` classes. Conversely, Rein [16] often misclassifies the background as `road` and trees as `building`. Furthermore, in the second clip, $tqdm$ shows better predictions especially for the `person` class, including the players and the spectators. These results highlight the promising generalization capabilities of our $tqdm$.

G. Comparison with Open-Vocabulary Segmentation

In this section, we compare our `tqdm` with Open-Vocabulary Segmentation (OVS) approaches [1, 7, 9, 17, 20], which also utilize language information from VLMs for segmentation tasks. The fundamental objective of OVS is to empower segmentation models to identify unseen classes during training. Our `tqdm` is designed to generalize across unseen domains for specific targeted classes, while OVS methods aim to segment unseen classes without emphasizing domain shift. This fundamental distinction leads to different philosophies in model design.




Method	Task	Backbone	G→C	G→B	G→M	Avg.
CAT-Seg [1] 	OVS	EVA02-L + Swin-B	57.30	51.24	61.83	56.79
<code>tqdm</code> (ours) 	DGSS	EVA02-L	68.88	59.18	70.10	66.05

Table S2: Comparison of mIoU (%; higher is better) with the state-of-the-art OVS method [1] trained on G and evaluated on C, B, M.  denotes EVA-02 [3, 4] pre-training. The best results are **highlighted** and our method is marked in **blue**.

We conduct a quantitative comparison with the state-of-the-art OVS method, namely CAT-Seg [1], on DGSS benchmarks. CAT-Seg optimizes the image-text similarity map via cost aggregation, and includes partial fine-tuning of the image encoder. For a fair comparison, both models utilize the EVA02-large backbone with EVA02-CLIP initialization and a 512×512 input crop size. As demonstrated in Tab. S2, `tqdm` outperforms the OVS method in DGSS benchmarks (*i.e.*, $G \rightarrow \{C, B, M\}$). We conclude that the two models exhibit different areas of specialization.

References

1. Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. arXiv preprint arXiv:2303.11797 (2023) [7](#)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) [3](#)
3. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023) [7](#)
4. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: Exploring the limits of masked visual representation learning at scale. In: CVPR (2023) [7](#)
5. Hümmer, C., Schwonberg, M., Zhong, L., Cao, H., Knoll, A., Gottschalk, H.: VLT-Seg: Simple transfer of CLIP-based vision-language representations for domain generalized semantic segmentation. arXiv preprint arXiv:2312.02021 (2023) [3](#)
6. Kim, S., Kim, D.h., Kim, H.: Texture learning domain randomization for domain generalized segmentation. ICCV (2023) [3](#), [4](#), [5](#), [6](#)
7. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022) [7](#)
8. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:2304.05653 (2023) [2](#)
9. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted CLIP. In: CVPR (2023) [7](#)
10. Neuhold, G., Ollmann, T., Rota Bulò, S., Kotschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017) [3](#)
11. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-aware domain generalized segmentation. In: CVPR (2022) [3](#)
12. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016) [2](#), [4](#), [6](#)
13. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016) [2](#), [3](#)
14. Sun, Q., Chen, H., Zheng, M., Wu, Z., Felsberg, M., Tang, Y.: IBAFormer: Intra-batch Attention Transformer for Domain Generalized Semantic Segmentation. arXiv preprint arXiv:2309.06282 (2023) [3](#)
15. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023) [3](#)
16. Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Lin, P., Wang, B., Chen, H., Zheng, J.: Stronger, fewer, & superior: Harnessing vision foundation models for domain generalized semantic segmentation. arXiv preprint arXiv:2312.04265 (2023) [4](#), [5](#), [6](#)
17. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: ECCV (2022) [7](#)
18. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100k: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020) [3](#)
19. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from CLIP. In: ECCV (2022) [2](#)

20. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: ZegCLIP: Towards adapting clip for zero-shot semantic segmentation. In: CVPR (2023) [7](#)