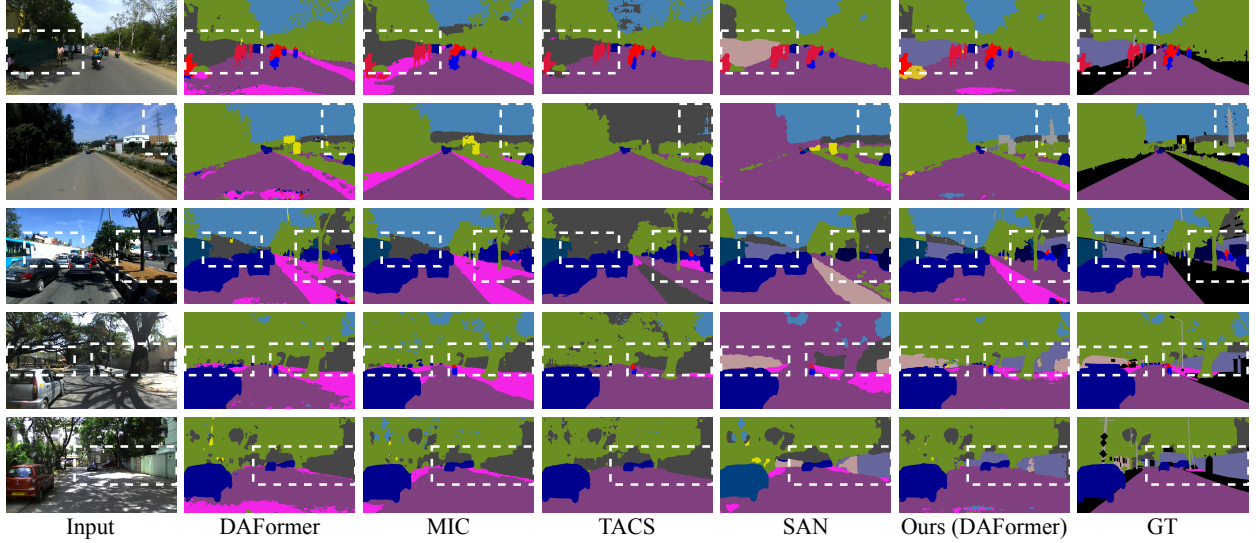


# Supplementary Materials: Cross-Class Domain Adaptive Semantic Segmentation with Visual Language Models

Anonymous Authors



**Figure 1: Segmentation results predicted by the previous SOTA UDA methods [3, 4], the VLM-based method [8], the CCDA method [2], and our presented framework on the SYNTHIA  $\rightarrow$  IDD benchmark. The novel classes are highlighted by white dashed boxes.**

## 1 OVERVIEW

In this supplementary material, we provide more information to support our proposed method. In Sec. 2, we extend the SOTA comparison on the SYNTHIA  $\rightarrow$  IDD benchmark and qualitatively discuss the generalizability of our proposed method with various UDA frameworks. In Sec. 3, we illustrate the potential limitations and promising future directions.

## 2 ADDITIONAL EXPERIMENTS

In our main paper, we validate the effectiveness and applicability of our method across various UDA frameworks and different CCDA scenarios. To further support the claims made in our main paper, here, we will provide qualitative results of various methods on the SYNTHIA  $\rightarrow$  IDD benchmark, as well as the performance gains of our method under different UDA frameworks.

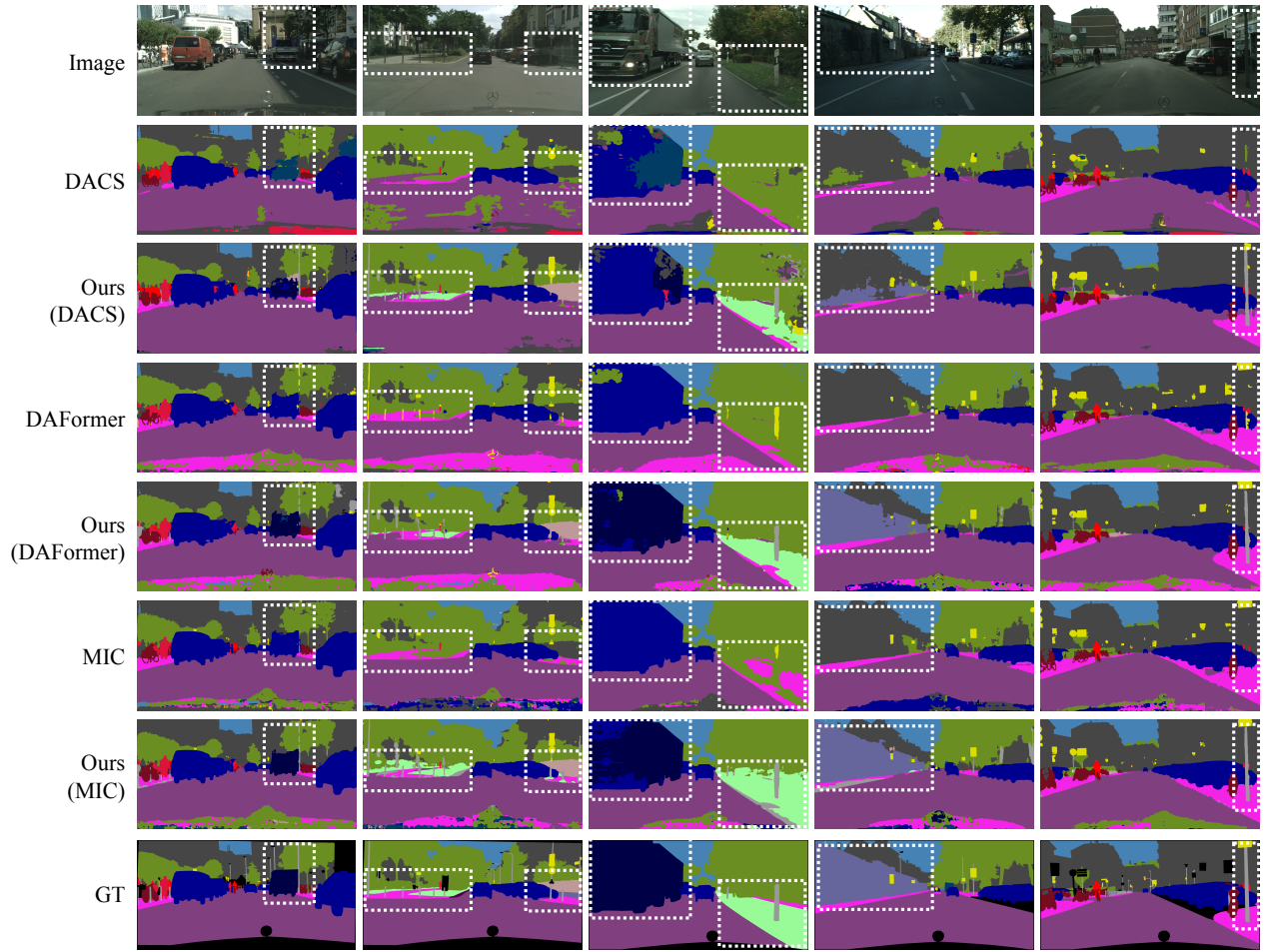
### 2.1 Comparison with SOTA methods

In our main paper, we only report the quantitative results of our proposed method and previous competitive methods [1–5] on the SYNTHIA  $\rightarrow$  IDD benchmark. Thus, in this section, we delve into a qualitative analysis of our proposed elements. As demonstrated in Figure 1, previous UDA methods [3, 4] often recognize novel classes as shared ones. In contrast, our proposed method addresses these issues by relabeling and augmenting pseudo labels with mask proposals for novel classes. For example, in the first row, both

DAFormer [3] and MIC [4] erroneously segment the “wall” as a “building”, due to their visual similarities. Instead, our proposed label alignment and novel class resampling methods provide numerous annotations for these novel classes, enabling our model to learn discriminative features from visually similar classes and consequently distinguish them from images. When comparing with the VLM-based method [8] and the CCDA method [2], our method also exhibits notable advantages in shared classes. Specifically, SAN [8] and TACS [2] tend to generate inaccurate predictions for shared classes, such as misidentifying the “sky” in the second row or the “trees” in the fourth row. However, our method effectively learns novel classes without compromising the segmentation performance on shared classes.

### 2.2 Generalization for various UDA methods

To further validate the robustness and generalizability of our method with various UDA frameworks [3, 4, 7], we present qualitative segmentation results in Figure 2. These results confirm that the positive impact brought by our method is not limited by UDA frameworks. Firstly, our approach demonstrates significant improvements in handling shared classes. For example, compared to DACS [7], our method notably contributes to restoring the complete visual shapes of shared classes like “road” in the second column and “sidewalk” in the last column. Secondly, our method excels in segmenting novel classes, as evidenced in the third column where “truck” and “terrain” are accurately distinguished from similar-looking shared



**Figure 2: Qualitative results of previous UDA methods [3, 4, 7] and our method with different UDA frameworks on the SYNTHIA → Cityscapes benchmark. The novel classes are highlighted by white dashed boxes.**

classes like “car” and “vegetation”, showcasing the robustness of our method in handling intricate semantic differences. These compelling findings collectively affirm the effectiveness of our proposed method and its commendable generalization capabilities across a spectrum of UDA frameworks.

### 3 LIMITATIONS AND FUTURE WORK

**Limitations.** In our CCDA setup, the novel classes in the target domain are predetermined. Relying on the names of these novel classes, we employ VLMs [1, 6] to localize the novel classes within unlabeled target images. However, in real-world scenarios, sometimes the names of novel classes are not known in advance. In this case, VLMs can not provide novel-class mask proposals for pseudo labels, leading to the limited applicability of our method in more challenging scenarios.

**Future work.** In our upcoming research, we aim to devise a method for achieving accurate segmentation of both shared and novel classes, even in cases where the names of the novel classes remain unknown. Specifically, developing a self-promoting approach

tailored for VLMs holds promise as a dependable solution to address the above limitation.

### REFERENCES

- [1] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.
- [2] Rui Gong, Martin Danelljan, Dengxin Dai, Danda Pani Paudel, Ajad Chhatkuli, Fisher Yu, and Luc Van Gool. 2022. TACS: Taxonomy adaptive cross-domain semantic segmentation. In *Proceedings of the European Conference on Computer Vision*. 19–35.
- [3] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9924–9935.
- [4] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11721–11732.
- [5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

233			
234	Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natu-	IEEE/CVF Winter Conference on Applications of Computer Vision. 1379–1389.	291
235	ral language supervision. In <i>Proceedings of the International Conference on Machine</i>	[8] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. 2023. Side	292
236	<i>Learning</i> . 8748–8763.	adapter network for open-vocabulary semantic segmentation. In <i>Proceedings of</i>	293
237	[7] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. 2021.	<i>the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> . 2945–2954.	294
238	DACS: Domain adaptation via cross-domain mixed sampling. In <i>Proceedings of the</i>		295
239			296
240			297
241			298
242			299
243			300
244			301
245			302
246			303
247			304
248			305
249			306
250			307
251			308
252			309
253			310
254			311
255			312
256			313
257			314
258			315
259			316
260			317
261			318
262			319
263			320
264			321
265			322
266			323
267			324
268			325
269			326
270			327
271			328
272			329
273			330
274			331
275			332
276			333
277			334
278			335
279			336
280			337
281			338
282			339
283			340
284			341
285			342
286			343
287			344
288			345
289			346
290			347
			348