Collaborating Foundation Models for Domain Generalized Semantic Segmentation

Supplementary Material

The supplementary material is organized as follows: We provide further details about the experiments in Sec. A, and in Sec. B we display a side-by-side qualitative comparison of CLOUDS and other state of the art methods. In Sec. C, we examine the qualitative effects of incorporating different foundation models into our framework. Sec. D shows the effect of SAM in refining the pseudo labels for self-training. Sec. E presents a collection of images produced by diffusion models, each guided by a distinct text prompt generated by the LLM.

A. Experimental details

Training losses: As mentionned in Sec. 3.1, the source loss $\mathcal{L}_S$ is a linear combination of the mask loss $\mathcal{L}_{\text{mask}}$ and classification loss $\mathcal{L}_{\text{cls}}$. The mask loss can be expressed as follows : $\mathcal{L}_{\text{mask}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}$, and we maintain the same values of λ_{ce}, λ_{dice} and λ_{dice} used in Mask2Former [9]. The self-training loss $\mathcal{L}_{ST}$ is exactly similar to $\mathcal{L}_S$, with the key difference being its application to the images generated by the model.

Prompting of the LLM: We used the Llama2-70b-Chat version of LLama2, which is optimized for dialogue use cases, for more details about the architecture and training details, refer to [68]. The full prompt that was used to prompt Llama2 is the following: *I want a list of prompts that can be used by an image generation model to generate synthetic images. The prompt should strictly follow this template: "a photo of X in Z" where X contains one or multiple class names within road, building, sidewalk, wall, fence, pole, traffic light, traffic sign, vegetation, grass, sky, person, rider, car, truck, bus, train, motorcycle, bicycle put in the context of an urban street scene, you use other synonyms to increase diversity. Z contains a brief description of regular lighting and weather conditions. Can you provide 100 diverse and simple prompts.*

B. Comparison with state of the art

The qualitative analysis in Figure A1 clearly shows the differences between CLOUDS and other approaches like GroundingSAM and SHADE. GroundingSAM struggles with fully segmenting all the regions due to its reliance on text prompts, which becomes challenging when objects are difficult to describe textually. Conversely, SHADE lacks robustness in distinguishing classes with similar features, such as "road" and "sidewalk", resulting in the creation of noisy pseudo labels.

C. Impact of each component

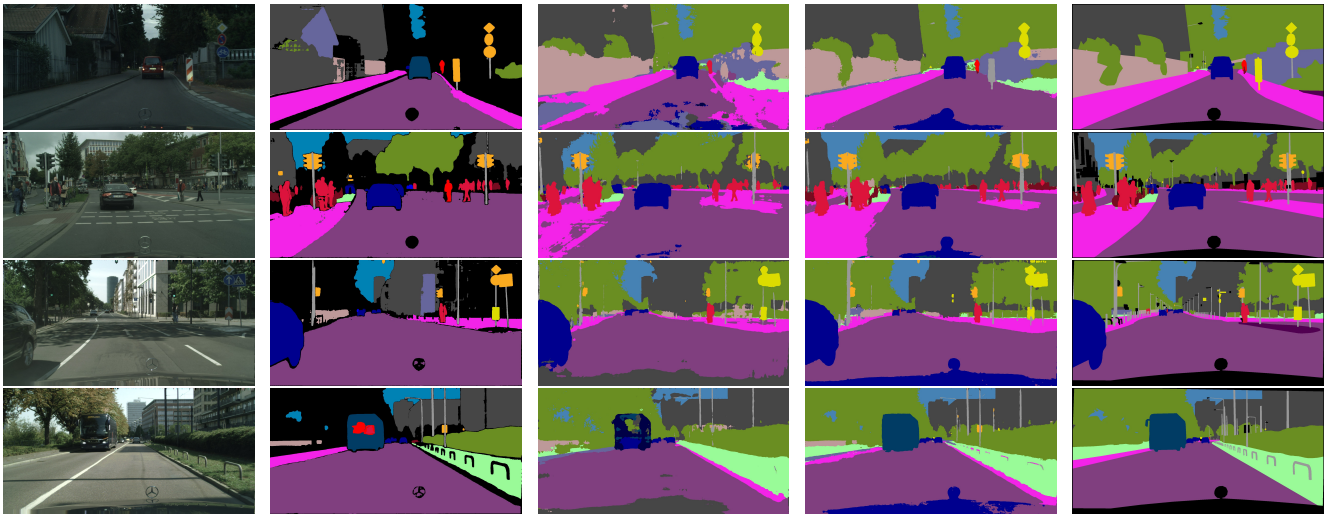
The qualitative analysis presented in Figure A2 illustrates the progressive impact of integrating various foundation models into our system. The initial employment of CLIP [56] as a sole feature extractor paired with Mask2Former [9] decoder yields a quantitative leap forward over previous DGSS methods, as shown in Table 4. However, this configuration, as visible in the second column of Figure A2, struggles to distinguish between similar classes. This issue is similarly observed in with SHADE [83], as shown in Figure A1. The incorporation of the {LLM [68] + Diffusion [59]} models with CLIP, showcased in the third column, improves segmentation quality and minimizes artifacts. This enhancement results from the introduction of a self-training loss, leveraging the original pseudo labels provided by the Teacher model. In the fourth column, we see the final enhancement: incorporating SAM to improve pseudo labels used in self-training. This leads to the distinctive sharp and detailed segmentation maps of our finalized model, CLOUDS. This step-by-step enhancement highlights the effectiveness of sequentially adding foundation models, each contributing to increasingly accurate and detailed segmentation outcomes.

D. Pseudo label refinement using SAM

In Figure A3, we show some examples of initial and refined pseudo-labels. This comparison reveals that the refined pseudo labels have more accurate object boundaries. The refinement process can effectively eliminate ambiguous objects, such as the "umbrella" depicted in the first row of Figure A3. However, this method is not without its limitations. Occasionally, it incorrectly designates some pixels as undefined (marked in black), excluding them from supervision despite their correct initial segmentation. It tends to discard numerous pixels, especially at the intersections of masks. Despite this, the pseudo label refinement reduces false positives within the pseudo labels.

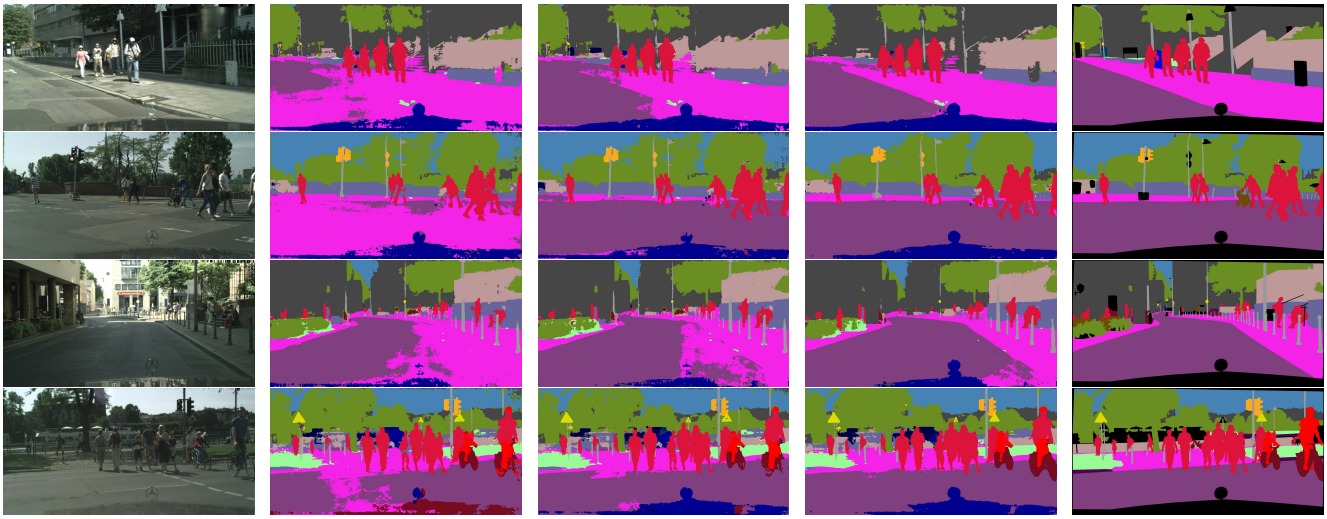
E. Generated images using the Diffusion model

In Figure A4, we show some generated images using the diffusion model along with their caption that were generated using LLM and used as text conditioning during generation.



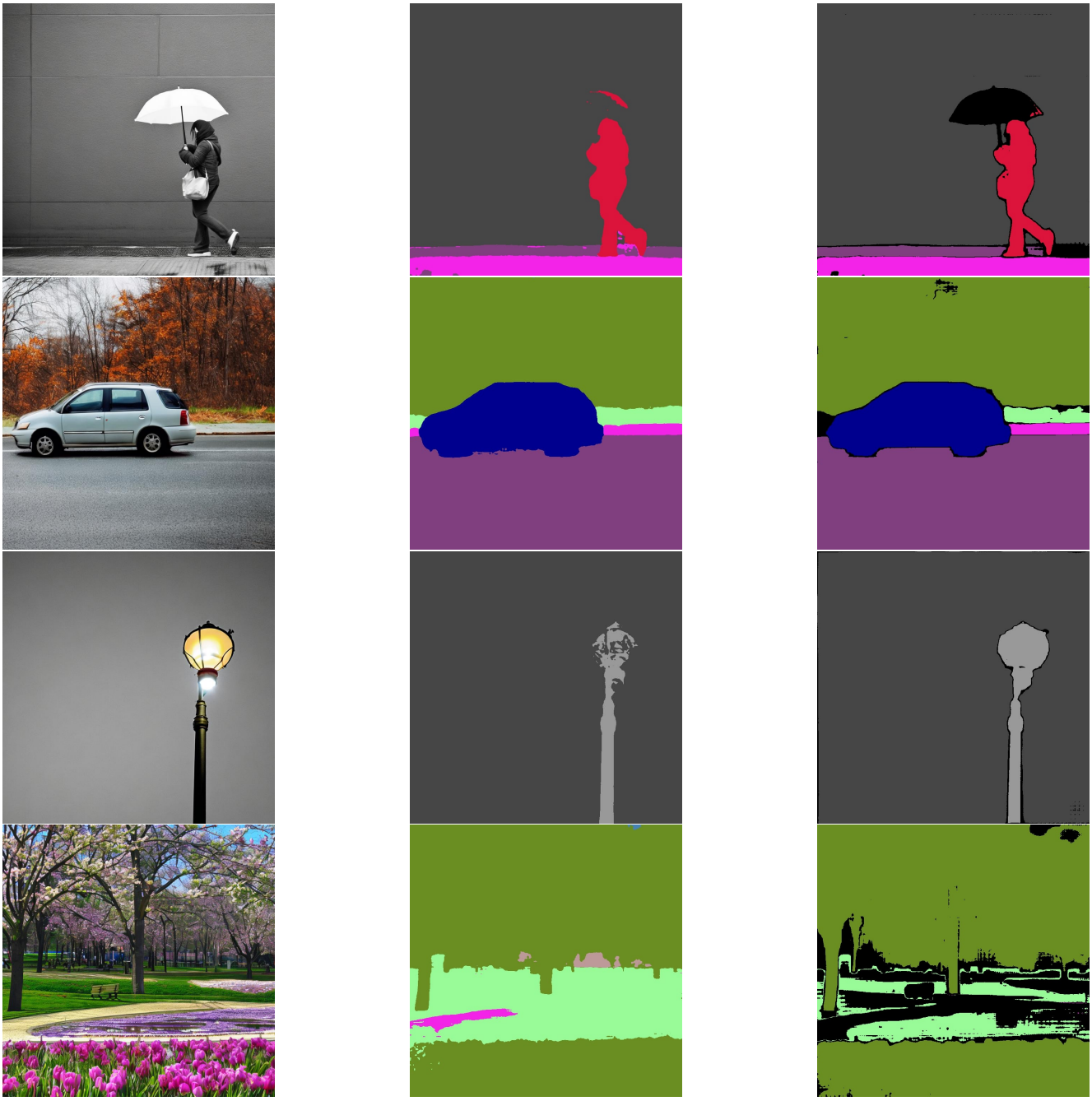
(a) RGB image (b) GroundingSAM [35, 41] (c) SHADE [83] (d) **CLOUDS** (e) GT

Figure A1. Qualitative study on the GTA $\rightarrow$ {C, B, M} scenario using ConvNext-L architecture. For each RGB image (a), we show (b) [35, 41] the segmentation map predicted by GroundingSAM [35, 41], (c) SHADE [83], (d) CLOUDS and (e) the Groundtruth associated.



(a) RGB image (b) CLIP (c) CLIP + {L, D} (d) CLIP + {L, D} + SAM (e) GT

Figure A2. Qualitative ablation study on the GTA $\rightarrow$ {C, B, M} scenario using ConvNext-L architecture. We demonstrate the impact of each foundation model in CLOUDS. For each RGB image (a), we show: (b) the segmentation maps predicted by the model using CLIP only as a feature extractor, (c) CLIP + {LLM+Diffusion} where self-training is done using the original pseudo labels, and (d) CLIP + {LLM+Diffusion} + SAM where the pseudo labels are refined using SAM



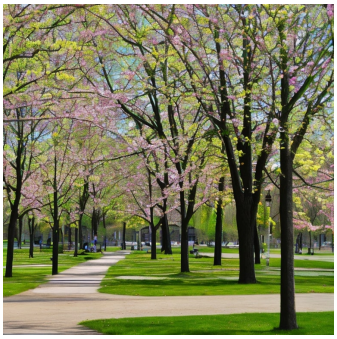
(a) RGB image (b) Pseudo label before SAM (c) Pseudo label after SAM

Figure A3. Pseudo label refinement using SAM. We show: (a) The generated image using the Diffusion model conditioned by the LLM, (b) the pseudo label before using SAM, and (c) the pseudo label after the refinement by SAM

(a) A photo of a busy road in daylight with sunny weather.

(b) A picture of a parked car on the side of the road on a cloudy day.

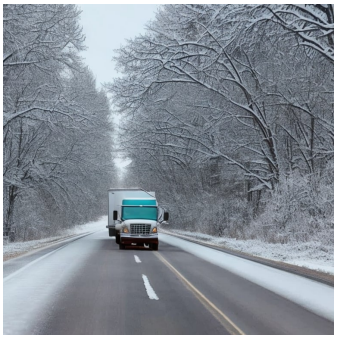
(c) A picture of a person on a motorcycle in the sun.



(d) A picture of a city park in the spring on a clear day

(e) A photo of a person sitting on a bench reading a book in the shade.

(f) A photo of a traffic light in the morning



(g) A snapshot of a truck driving down the road in the winter

(h) A photo of a traffic sign in the shape of a yield sign in the snow

(i) A picture of a car parked on the side of the road in the fall

(j) A photo of a person riding a bicycle in the park on a sunny day

(k) A snapshot of a truck driving down a highway in the night

(l) A photo of a group of pedestrians crossing the street in the snow

Figure A4. Examples of generated images using the diffusion model along with the text conditioning prompt generated by the LLM