Semantic Segmentation using Foundation Models for Cultural Heritage: an Experimental Study on Notre-Dame de Paris

Kévin Réby, Anaïs Guillem, Livio De Luca

Semantic Segmentation using Foundation Models for Cultural Heritage: an Experimental Study on Notre-Dame de Paris

Kévin Réby, Anaïs Guillem, Livio De Luca

UMR CNRS/MC 3495 MAP, 13402 Marseille, France

Abstract

Vision foundation models have already had a major impact on several computer vision tasks. This work aims to study their usefulness in the context of cultural heritage. By utilizing the Segment Anything Model (SAM) we could perform segmentation on Notre-Dame de Paris images. Additionally, we have developed a pipeline that combines various foundation models (GroundingDINO and CLIP) to demonstrate their abilities for semantic segmentation of cultural heritage data.

Methods

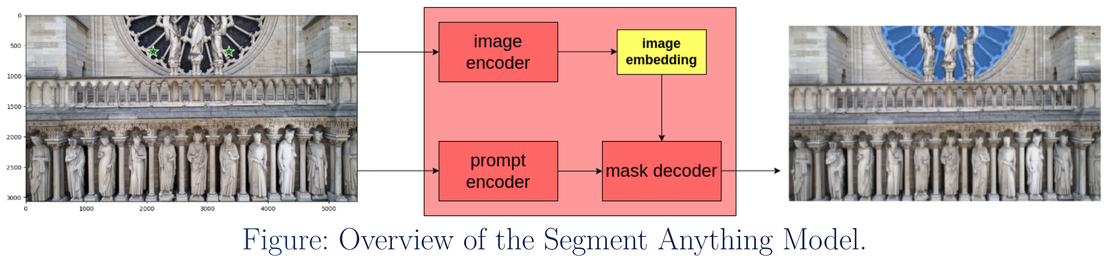

Figure: Overview of the Segment Anything Model.

Foundation models have already had a major impact on Natural Language Processing, but in computer vision, pre-trained models on labeled datasets like ImageNet are still standard practice. To our knowledge, foundation models have never been used in the context of cultural heritage. The presented experiment uses foundation models to build a pipeline capable of automatically: (1) segmenting images and (2) labeling them with concepts that are described in our controlled vocabularies. Our method combines several foundation models to perform semantic segmentation:

- First, we use Grounding DINO[3] to detect objects and get bounding boxes since we have observed that SAM performs best with these kinds of prompt inputs.
- Then we pass them into SAM[2] to automatically segment objects in images
- Finally, we use CLIP[1] to label the segmentation masks obtained from the previous step. The tags for this last step come from our documented domain-specific thesauri.

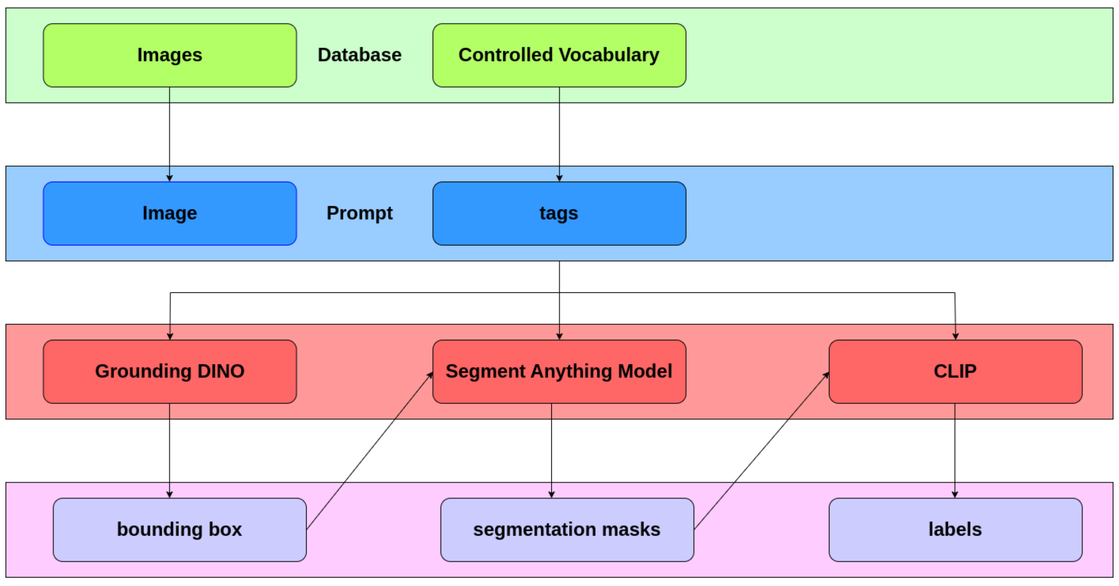

Figure: Our pipeline for semantic segmentation using foundation models.

Acknowledgements

This work was supported by the CNRS and the French Ministry of Culture within the European Research Council Advanced Grant 'nDame_Heritage: n-Dimensional analysis and memorization ecosystem for building cathedrals of knowledge in Heritage Science'. The authors also wish to thank AGP who shared their acquisition campaigns, and all the scientific partners and collaborators working on Notre-Dame de Paris.

Results

We compare SAM with zero-shot segmentation models from the state of the art. Our results show that SAM outperforms other segmentation models, especially when the image shows an unusual view of the cathedral.

Models	SAM	OpenSeed	CLIPSeg
mIoU	77.0	35.3	48.32

Table: Comparison of zero-shot segmentation models on our dataset.

We also compare the prediction output by SAM given different prompts: no prompt, points, and bounding boxes. We observe that SAM performs best when bounding boxes are used as prompt inputs to guide the segmentation. We have also studied the influence of various text prompts in our pipeline. It performs slightly better with selected terms when experts have selected the right words to generate the best output.

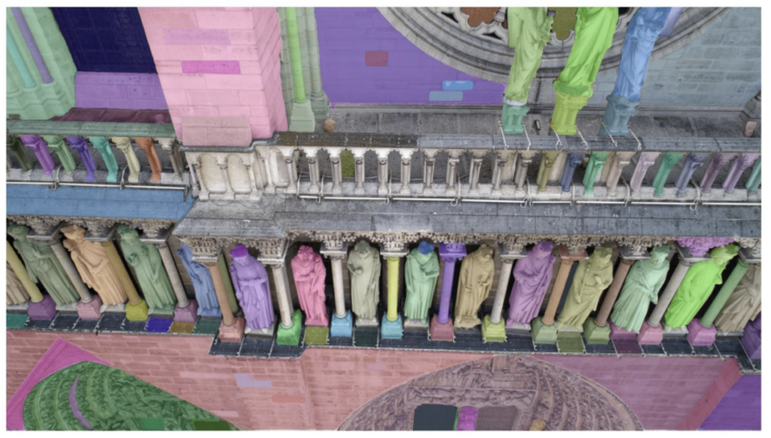

Figure: Example of automatic segmentation using SAM.

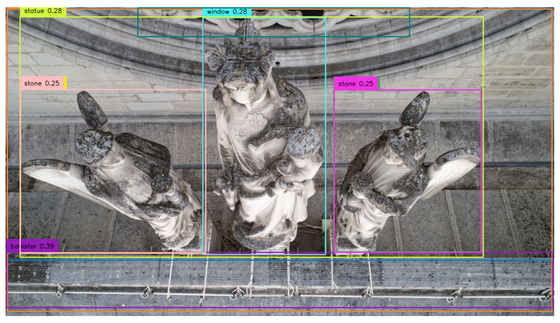

Figure: Example of object detection using Grounding DINO.

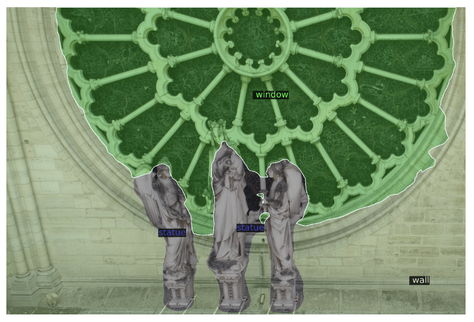

Figure: Example of semantic segmentation using foundation models.

Conclusion

We have demonstrated the capabilities of foundation models for zero-shot semantic segmentation on cultural heritage data, using our Notre-Dame de Paris dataset as an example. We were able to perform segmentation on the Notre-Dame images. However, our current evaluation is limited to certain sections of the Cathedral. For future work, we plan to further test our pipeline on other parts of the cathedral, with a particular focus on the interior.

References

[1] Alec Radford et al. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[2] Alexander Kirillov et al. Segment anything. *arXiv:2304.02643*, 2023.

[3] Liu Shilong et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023.