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

Kévin Réby, Anaïs Guilhelm, Livio De Luca
UMR CNRS/MC 3495 MAP
13402 Marseille, France
Email: firstname.surname@map.cnrs.fr

Abstract

The zero-shot performance of foundation models has captured a lot of attention. Specifically, the Segment Anything Model (SAM) has gained popularity in computer vision due to its label-free segmentation capabilities. Our study proposes using SAM on cultural heritage data, specifically images of Notre-Dame de Paris, with a controlled vocabulary. SAM can successfully identify objects within the cathedral. To further improve segmentation, we utilized Grounding DINO to detect objects and CLIP to automatically add labels from the segmentation masks generated by SAM. Our study demonstrates the usefulness of foundation models for zero-shot semantic segmentation of cultural heritage data.

1. Introduction

On April 15, 2019, a fire destroyed part of Notre Dame Cathedral in Paris. The fire broke out in the frame and lasted for almost 15 hours. The 400 mobilized firefighters managed to extinguish the fire early in the morning after several hours of firefighting. The fire is believed to have started in the attic at the foot of the tower overlooking the cathedral's transept. The fire destroyed the roof and the 93-meter-tall tower, causing it to collapse and part of the vaults and nave to be lost. [26]. It was rebuilt in the 19th century by Viollet-le-Duc and was currently undergoing restoration. The roof of the cathedral was also reduced to rubble. It was the largest and most serious fire since the cathedral was built in 1163. As soon as the cathedral caught fire, many people tried to help with the restoration work. To organize the large-scale research work that followed, the French Ministry of Culture commissioned the 'National Center for Scientific Research' (CNRS) to lead the scientific effort to enable new discoveries and restoration work. The cathedral was declared a UNESCO World Heritage Site in 1991.

Due to the importance of Notre Dame as a historical monument, its destruction has been widely mediated. It soon became an unprecedented and unique opportunity for the scientific community to study its restoration. A digital data working group has coordinated scientific projects enabling digital data management from scientific research and restoration work[1]. The European Research Council (ERC) Advanced Grant project 'nDame Heritage' combines the digital humanities with computer science and artificial intelligence to create a collaborative knowledge system that analyzes multiple views of the same heritage by different experts. The aim is to describe the scientific activities of the 175 researchers in order to understand the process of scientific knowledge generation and its dissemination[9, 10][2].

In this paper, we combine digital humanities with foundation models in order to process, integrate, and enrich Notre-Dame's data[9]. Our contributions can be summarized as followed:

- we build a dataset from Notre-Dame data, with 2d images and a controlled vocabulary,
- we compare SAM with several segmentation models,
- we test the efficacy of several prompts on SAM,
- we build a pipeline using several foundation models for semantic segmentation on cultural heritage data.

2. Related work

2.1. Foundation models in Computer Vision

Foundation models. Over the past few years, numerous applications of vision, robotics, Natural Language Processing (NLP), and the broader deep learning community have advanced significantly thanks to the potential

[1] https://www.notre-dame.science/
[2] http://www.ndameheritage.map.cnrs.fr/

of large-scale multimodal data and large models. A 'foundation model' (a term popularized by the Stanford Institute for Human-Centered Artificial Intelligence[22]) is essentially a large deep learning model that has been pre-trained on a massive amount of data. For example, Generative Pre-trained Transformer 3 (GPT-3) has been trained on about 45 TB text data from multiple sources (including Wikipedia) and has 175 billion parameters[23]. In NLP, Large language models (LLMs) like GPT-3[23] and BERT (Bidirectional Encoder Representations from Transformers)[12] have achieved cutting-edge outcomes in a variety of NLP applications and have since expanded into other fields such as Computer Vision (Vision Transformers (VIT)[14], Stable Diffusion[46]) and Audio (Whisper[43], XLS-R[13]). LLMs are trained on huge datasets and have impressive performance compared to classic fine-tuned models in few/zero-shot generalization[23]. These models can generalize tasks beyond those seen during training, using prompt engineering to generate a valid text as an output. A foundation model is designed to be adapted or fine-tuned to a variety of downstream tasks. Without being expressly trained on them, foundation models can accomplish a variety of functions. For example, given a short natural language prompt, foundation models such as DALLE 2[45] or Midjourney[4], may execute tasks like answering questions, producing text, or generating images.

Visual Foundation models. Foundation models, such as BERT[12] and GPT-4[41], have already had a major impact on NLP, but in computer vision, pre-trained models on labeled datasets like ImageNet[11] is still standard practice. Foundation models are mostly based on the transformer architecture[17]. Given the success of transformers in NLP, researchers also try to apply them in computer vision tasks. A lot of papers have used different versions of BERT architecture: VisualBERT[33], ViLBERT[36], Pixel-BERT[28], VL-BERT[50], etc. Recently, large-scale pre-training methods have been developed that learn directly from web-scale images, such as CLIP (Contrastive Language-Image Pre-Training)[15], ALIGN (A Large-scale ImaGe and Noisy-Text Embedding)[18]. Text pairs show very encouraging progress in efficient transfer learning and zero-shot capabilities. However, such models are limited to image-to-text mapping tasks such as classification, searching, etc. Foundation models are made possible by transfer learning and scaling. Transfer learning is the process of applying "knowledge" from one task (for example, object detection in photographs) to another (for example, activity recognition in films)[59]. In deep learning, pretraining is the most common way of transfer learning: a model is trained on a surrogate task and then modified to meet the downstream task of interest. For example, in computer vision, pretraining on the ImageNet dataset for image classification is a well-known approach to transfer learning using this annotated dataset that has been around for at least a decade[11]. Self-supervised learning, on the other hand, generates the pretraining problem automatically from unannotated data[1]. For example, the masked language modeling challenge used to train BERT asks participants to predict the missing word in a sentence given the context of the rest of the sentence[12]. Another example in computer vision is DINO ('DIstillation with NO labels'), a self-supervised trained model for classification or segmentation tasks[6]. DINO is a foundation model that aims to learn effective visual representations without the need for manually labeled data. This model is able to learn effective visual representations without the need for manually labeled data. It is based on the ViT (Vision Transformer) architecture to learn and represent visual information. DINO uses a teacher-student framework where the teacher network, built with a momentum encoder, predicts the output, and the student network tries to replicate it. The authors used a self-distillation approach, where the student model learns from the teacher's predictions through a standard cross-entropy loss. By aligning the student's representations with the teacher's, the model helps the student model improve its understanding of visual features and semantic layout. DINOv2 is its improved version[21].

Multimodal Foundation models. Recently, numerous studies have begun to use the synchronization or alignment of many modalities (audio-visual/visual-language correspondences, spatio-temporal image sequences with associated optical flow, camera motion, etc.) as self-supervision to extract knowledge from vast amounts of unlabeled data. Exploiting the cross-modal connection between vision and language (text) for data generation is one of the well-known examples, as shown by models like Stable Diffusion[46], DALL-E[16], ImaGen[19] and others. Models like CLIP[15] or ALIGN[18] are based on contrastive learning. Although transfer learning makes foundation models practical, their power lies from their size[3]. The development of Transformer-based model architecture, which takes advantage of the parallelism of the hardware to train models that are much more expressive than before, and finally the availability of much more training data enabled the rise of foundation models. For example, training the Segment Anything Model (SAM) from Meta 256 GPUs requires 11 million images and 1 billion segmentation masks[31].

2.2. Foundation models for Segmentation

These recent developments highlight the growing focus on foundation models for segmentation and their potential applications in various domains, such as general image segmentation and medical image segmentation.

One of the first models viewed as a foundation model for computer vision by its authors was Florence[20]. Florence can be adapted for various computer vision tasks, such as classification, action recognition, object detection, Visual Question Answering (VQA), image captioning, etc. In the medical field, a foundation model called UniverSeg has been developed for medical image segmentation. An empirical analysis was conducted comparing this model to the conventional approach of training a task-specific segmentation model. The study focused on prostate imaging and evaluated the performance of UniverSeg in this context[30]. Other foundation models were proposed for endoscopy video analysis[54] and 3d medical image segmentation[52].

SegGPT is another generalist model for segmenting everything in context[53]. This model is built on the GPT ('Generative Pre-trained Transformer') model, a language model that has been successful in many natural language processing tasks. The authors applied a segmentation mask to the last layer of the GPT model, leveraging its ability to generate coherent and structured sequences of output tokens to produce accurate segmentation results in its specific context. The SegGPT model can be trained in a supervised or unsupervised learning regime using various loss functions. The supervised training involves optimizing pixel-wise classification losses, while unsupervised training focuses on context-aware coloring tasks. SegGPT is an example of how the GPT model architecture can be used beyond its original purpose of natural language processing and applied to computer vision tasks.

2.3. Segmentation using deep learning for Cultural Heritage

In this project, we want to use semantic segmentation methods in order to help experts in annotating low-level objects like statues in our 2d images' corpus. With the successful application of deep learning methods in computer vision, several models have been developed for semantic segmentation. In 2015, Long et al. proposed fully convolutional networks (FCNs). The important idea of FCN is to use convolution instead of full connection, which made it possible to input any image size[35]. Convolutional neural networks (CNNs) achieved good results in image classification, whose output layers are the categories of images[51].

However, semantic segmentation needs to map the high-level features back to the original image size after obtaining high-level semantic information. This requires an encoder–decoder architecture. For example, the U-Net architecture became very popular for semantic segmentation. Originally developed for medical images[47], U-Net had great success in other fields like agriculture[57] or satellite images[29]. SegNet[2] is also an Encoder-Decoder network based on VGG-16, a standard CNN architecture with 16 convolutional layers. DeepLab[7] improves FCN by employing atrous convolution.

Recent news in the field of AI for cultural heritage includes the use of machine and deep learning to aid in the restoration and preservation of cultural heritage[24]. Deep learning was used for semantic segmentation in cultural heritage, mostly using 3d points cloud[55, 5]. Perdicca et al.[42] built a model called DGCNN (Dynamic Graph Convolutional Neural Network) for point cloud segmentation on the ArCH (Architectural Cultural Heritage) dataset. Matrone et al. proposed an architecture named DGCNN-Mod+3Dfeat that combines both methodologies' positive aspects and advantages for 3D cultural semantic segmentation in point clouds[39]. Croce et al. used CNN (Convolutional Neural Network)[8] trained on the Camvid dataset for 2d image classification. These recent developments showcase the growing interest and the potential for deep learning and foundation models for cultural heritage restoration and preservation.

To our knowledge, foundation models have never been used in the context of cultural heritage. In this work, our objective is to use them to perform semantic segmentation on the Notre Dame dataset.

3. Semantic segmentation using Foundation models

As explained before, the capabilities of foundation models have been tested in various fields, including segmentation. Multimodal foundation models, which are simultaneously trained using many modalities, have demonstrated great effectiveness in various applications, including cross-modal retrieval, zero-shot categorization, text-to-image/video/3D production, and image segmentation. The presented experiment uses foundation models to construct a pipeline capable of automatically: (1) segmenting images and (2) labeling them with concepts that are described in our controlled vocabularies.

3.1. The Notre-Dame dataset

The Notre-Dame de Paris scientific research team, which currently includes 175 researchers from archaeology, anthropology, architecture, history, chemistry, physics, and computer science, is the ideal experimental setting for establishing a new domain of interdisciplinary and multidimensional data as the starting point for studying knowledge generation processes in cultural assets. We are creating a data corpus that is representative of the academic practice of contemporary cultural heritage research. As a result of this unique opportunity to collect and analyze large amounts of scientific data, 'nDame Heritage' aims to provide a generalizable approach, reproducible methodology, and an open and reusable digital environment through collaborative research. The objective is

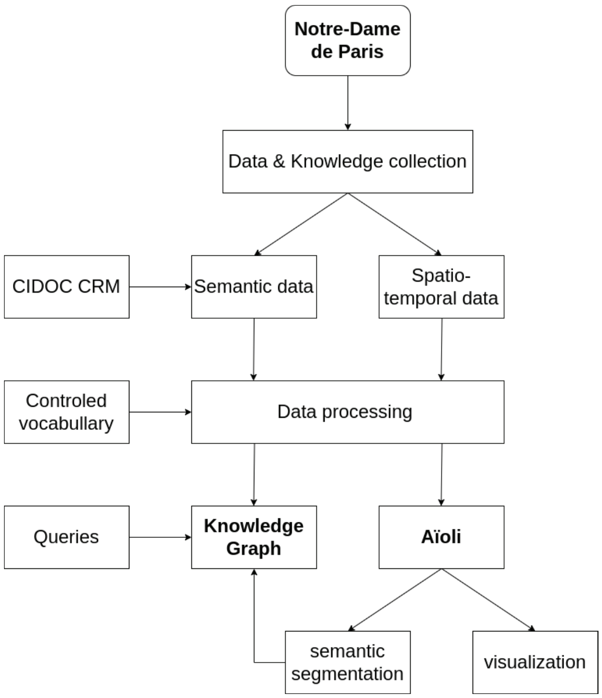

Figure 1. Overview of the nDame Heritage knowledge ecosystem.

to build a knowledge system in order to use all the data generated by many scientists and to create an environment for spatial, temporal, and semantic analysis in cultural heritage research. This empirical data integration approach generates new data analysis perspectives to explore how each discipline is connected to each other's work, within research activities by multiple actors.

We build a semantically enriched corpus of data, by introducing a groundbreaking approach through the cross-section of a data enrichment process, capable of describing digital assets not only through conventional metadata but also by storing the different steps that scientists take on their way from raw data to interpretation and knowledge.

Since the beginning of the project, a lot of data types produced by Notre Dame Scientific projects, from various disciplinary profiles, have been included in the data corpus: bibliography, material sampling analysis (from multiple physicochemical characterization processes), technical surveys, drawings, photographs, NDT (Non-Destructive Techniques) imaging, mechanical and acoustic simulations, press and web resources, interviews, video documentaries, etc. Data already collected (and partially documented) from cultural institutions, research laboratories, and private companies include:

- 180 000 photographs (before and after the fire, during the restoration);
- 5000 3D point clouds (before and after the fire, during the restoration);
- hundreds of technical drawings (before the fire);
- dozens of structured 3D models relating to the cathedral's condition before and after the fire;
- 5000 documentary sources (archives, bibliography, iconography) relating to the cathedral's history.

Our digital environment (see Fig. 1) integrated a platform that now provides dedicated web services to more than 100 registered expert users for the management of deposits, the indexing of multimedia content, the structuring of thesauri and the CIDOC-CRM ontology [3]), the visualization of 3D digitization of the remains, the annotation of 2D images and 3D point clouds with semantic tags and the analysis of 4D datasets merging past-present-future states of the cathedral[10]. Using this ecosystem, we gathered collections of images from the 276 photogrammetric scenes covering the different parts of the cathedral. Thanks to the annotation ad visualization software 'Aïoli' already developed by the the MAP-CNRS team[38, 48], the researchers' observations are annotated in 2D images and propagated into 3D scenes. Nevertheless, these experts' annotations missed low-level characterization of building objects (*e.g.* stone, rose window, column, etc.). Therefore, we decided to use vision foundation models for automatic semantic segmentation. For this task, we have selected a part of our dataset that only concerns different views of the front of the cathedral.

3.2. Segmentation using SAM

Since the experts only focus on high-level annotations, we decided to use foundation models for automatic semantic segmentation and labeling of 2d images to assist specialists in annotating low-level objects (*e.g.* stone, rose window, column, etc.). The objective is to free researchers from this very time-consuming task, so they can focus on their subject of expertise.

One of the recent developments in foundation models for segmentation is the release of the Segment Anything Model (SAM) and the Segment Anything 1-Billion mask dataset (SA-1B) by Meta, making them available to the research community[4]. SAM is a vision foundation model for image segmentation and zero-shot learning. This model is designed to facilitate the segmentation of any object from any image, using prompts. SAM returns a segmentation mask when given a prompt, which can be a set of points, bounding boxes, or text[31]. The generalization capabilities of SAM were tested on several various tasks like medical images[40, 27, 49] or crater detection[25]. SAM is expected to accelerate computer vision research and foster advancements in segmentation tasks. This foundation model aims to

[3] https://www.cidoc-crm.org/

[4] https://segment-anything.com/

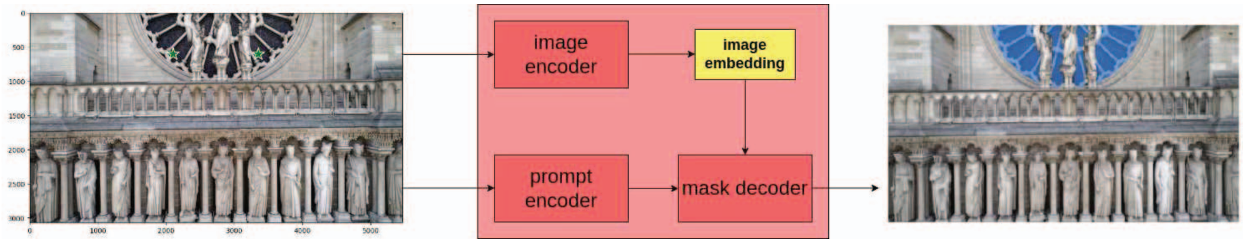

Figure 2. Overview of the Segment Anything Model.

Figure 3. Examples of automatic segmentation using SAM.

enable a wide range of applications and further research into image segmentation. SAM was already adapted for various tasks like image matting[32], point tracking[44], or image tagging[58]. Our goal is to use it for cultural heritage data.

As shown in Fig.3, SAM is composed of three modules: (1) an image encoder, (2) a prompt encoder, and (3) a mask decoder. The image encoder is a MaskedAutoEncoder (MAE) based on a ViT model and is used for image feature extraction. It takes an image of resolution 1024×1024 and generates an image embedding of size 64×64. The prompt encoder takes the positional information from the prompt input *i.e*. a set of points and/or bounding boxes to provide additional information for the decoder. The mask decoder is a two-layer transformer-based model for final mask predictions. However, SAM lacks the capability to output semantic labels associated with the segmentation masks.

We have tested the zero-shot segmentation capacity of SAM on our dataset. Zero-shot learning is a way to generalize on unseen labels, without having specifically trained to classify them. We compare SAM with zero-shot segmentation models from the state of the art, OpenSeed[60, 56] and CLIPSeg[37], using the mean Intersection-Over-Union (mIoU) metric. The IoU is defined as the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth. We did not train any models but used the publicly available provided weights and model configurations. The evaluation was conducted on an NVIDIA RTX A4500. All models just use an image and no additional prompt as input. Our results (see Table 1) show that SAM outperforms other segmentation models, especially when the image shows an unusual view of the cathedral (*e.g*. from above, see left picture on Fig. 3 and Fig. 5).

Models	SAM	OpenSeed	CLIPSeg
mIoU	77.0	35.3	48.32

Table 1. Comparison of zero-shot segmentation models on our dataset.

3.3. Qualitative analysis

We qualitatively compare the output of the prediction by SAM given different prompts. This task can be viewed as prompt engineering. Originally, prompt engineering is the process of enhancing the output of large language models (LLMs) like ChatGPT. It involves using crafted inputs known as prompts that guide the model in generating high-quality and relevant output. Therefore, prompt engineering requires a good understanding of both data and models. In our case, we have tested several types of prompt inputs in addition to the images: 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. First, we tested no text, then selected terms from our controlled vocabulary (like window, column, banister, statue, etc.), all the terms that concern the whole dataset, and finally, all vocabulary with some extra terms (*i.e*. terms that we know they are not present in our dataset) to see if it perturbs the model. We observe SAM performs slightly better with selected terms when experts have selected the right words to generate the best output.

3.4. Semantic segmentation with foundation models

Our pipeline combines several foundation models to perform semantic segmentation, as described in Fig.4. We can divide this process into three steps: first, we use Grounding DINO to detect object and get bounding boxes, since we have observed that SAM perform best with these kinds of prompt inputs. Then we pass them into SAM to automatically segment objects in images, and finally, we use CLIP to label the segmentation masks obtained from the previous step. The tags for this last step come from our documented domain-specific thesauri.

Since our results have shown that SAM performs best with bounding boxes as inputs, we use Grounding DINO to automatically detect objects in images. Grounding DINO [34] is a zero-shot object detector. Its architecture comprises an image backbone that extracts image features, a text backbone for text features, a feature enhancer for combining these features, a language-guided module for query selection, and finally a cross-modality decoder for refining bounding boxes.

These images with bounding boxes (see Fig. 5) are then used as inputs for SAM, with the controlled vocabulary, to perform the automatic segmentation as we explained before. Nevertheless, SAM does not have the capability to output semantic labels associated with the segmentation masks. Therefore, we decide to use another vision language foundation model: CLIP[15]. CLIP stands for 'Contrastive Language–Image Pre-training' developed by OpenAI in 2021[5]. It is a zero-shot model: given an image and text descriptions, the model can predict the most relevant text description for that image, without optimizing for a particular task. CLIP learns image representations from natural language supervision. It jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. CLIP was trained using 400 million images-text pairs. In the end, by combining those foundation models we obtain semantic masks with labels associated, like in Fig. 6.

[5]https://openai.com/research/clip

4. Conclusion and discussion

We have demonstrated the capabilities of foundation models for zero-shot semantic segmentation on cultural heritage data, using the Notre-Dame dataset as an example. By utilizing the Segment Anything Model (SAM), a vision foundation model, we were able to perform segmentation on the Notre-Dame dataset. Additionally, we have developed a pipeline that combines various foundation models to demonstrate their abilities for semantic segmentation. However, our current evaluation is limited to certain sections of the Notre-Dame de Paris Cathedral. For future work, we plan to further test our pipeline on other parts of the cathedral, with a particular focus on the interior.

Acknowledgments

This work was supported by the CNRS and the French Ministry of Culture within by the European Research Council (ERC) 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.

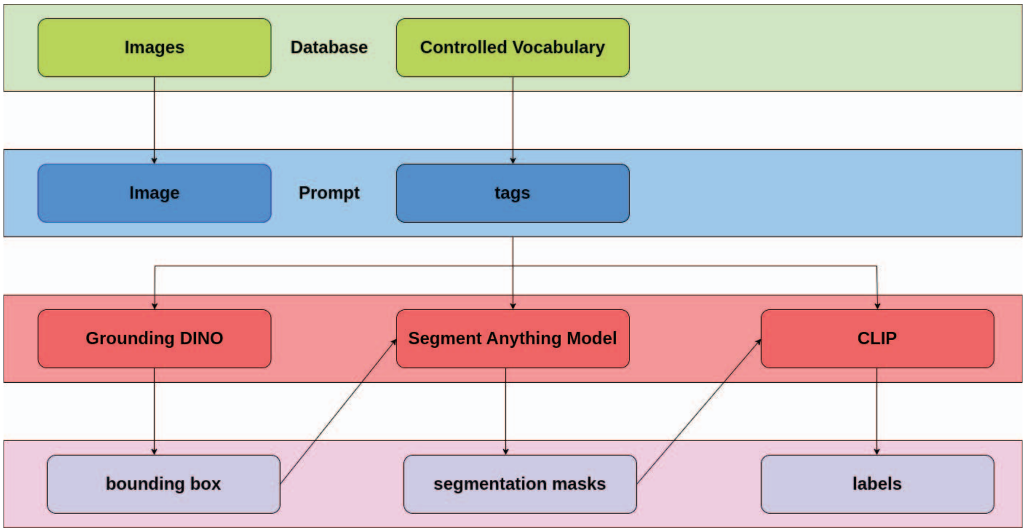

Figure 4. Our pipeline for semantic segmentation using foundation models.

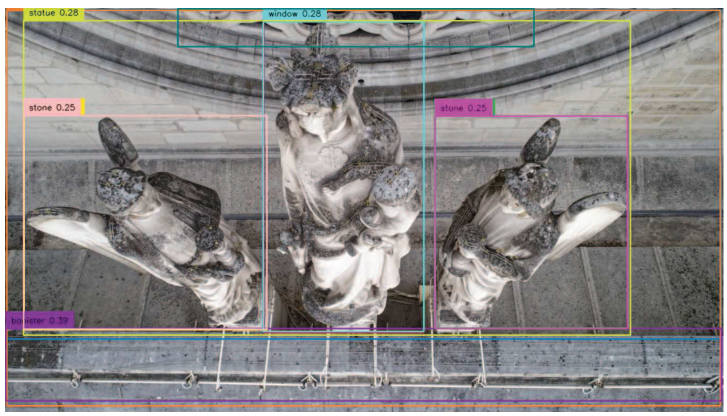

Figure 5. Example of object detection using Grounding DINO.

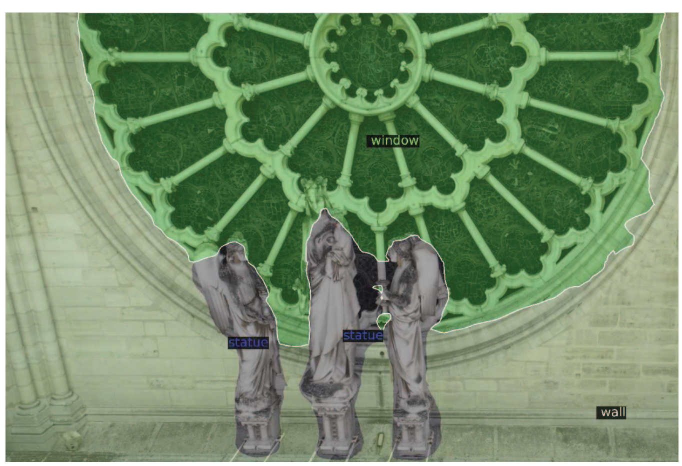

Figure 6. Example of semantic segmentation using foundation models.

References

[1] Mahmoud et al. Assran. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 2

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 3

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[4] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2, 2023. 2

[5] Yuwei Cao, Simone Teruggi, Francesco Fassi, and Marco Scaioni. A comprehensive understanding of machine learning and deep learning methods for 3d architectural cultural heritage point cloud semantic segmentation. In *Italian Conference on Geomatics and Geospatial Technologies*, pages 329–341. Springer, 2022. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[8] Valeria Croce, Adeline Manuel, Gabriella Caroti, Andrea Piemonte, Livio De Luca, and Philippe Véron. Semi-automatic classification of digital heritage on the aïoli open source 2d/3d annotation platform via machine learning and deep learning. *Journal of Cultural Heritage*, 62:187–197, 2023. 3

[9] Livio De Luca. Towards the semantic-aware 3d digitisation of architectural heritage: The" notre-dame de paris" digital twin project. In *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 3–4, 2020. 1

[10] Livio De Luca, Violette Abergel, Anaïs Guillem, Olivier Malavergne, Adeline Manuel, Ariane Néroulidis, Roxane Roussel, Miled Rousset, and Sarah Tournon. L'écosystème numérique n-dame pour l'analyse et la mémorisation multi-dimensionnelle du chantier scientifique Notre-Dame-de-Paris. SCAN'22 - 10e Séminaire de Conception Architecturale Numérique, Oct. 2022. Poster. 1, 4

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 2

[13] Arun Babu et al. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296, 2021. 2

[14] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[15] Alec Radford et al. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2, 6

[16] Aditya Ramesh et al. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. 2

[17] Ashish Vaswani et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2

[18] Chao Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021. 2

[19] Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[20] Lu Yuan et al. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 3

[21] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision, 2023. 2

[22] Rishi Bommasani et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. 2

[23] Tom B. Brown et al. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. 2

[24] Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108, 2020. 3

[25] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. Deep learning universal crater detection using segment anything model (sam). *arXiv preprint arXiv:2304.07764*, 2023. 4

[26] Antoine Gros, Anaïs Guillem, Livio De Luca, Élise Baillieul, Benoit Duvocelle, Olivier Malavergne, Lise Leroux, and Thierry Zimmer. Faceting the post-disaster built heritage reconstruction process within the digital twin framework for notre-dame de paris. *Scientific Reports*, 13(1):5981, 2023. 1

[27] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023. 4

[28] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[29] David John and Ce Zhang. An attention-based u-net for detecting deforestation within satellite sensor imagery. *International Journal of Applied Earth Observation and Geoinformation*, 107:102685, 2022. 3

[30] Heejong Kim, Victor Ion Butoi, Adrian V. Dalca, and Mert R. Sabuncu. Empirical analysis of a segmentation foundation model in prostate imaging, 2023. 3

[31] journal=arXiv preprint arXiv:2304.02643 year=2023 Kirillov, Alexander et al. Segment anything. 2, 4

[32] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. *arXiv preprint arXiv:2306.05399*, 2023. 5

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[37] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 5

[38] Adeline Manuel and Violette Abergel. Aïoli, a reality-based annotation cloud platform for the collaborative documentation of cultural heritage artefacts. In *Un patrimoine pour l'avenir, une science pour le patrimoine*, 2022. 4

[39] Francesca et al. Matrone. Comparing machine and deep learning methods for large 3d heritage semantic segmentation. *ISPRS International Journal of Geo-Information*, 9(9), 2020. 3

[40] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *arXiv preprint arXiv:2304.10517*, 2023. 4

[41] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2

[42] Roberto et al. Pierdicca. Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sensing*, 12(6):1005, 2020. 3

[43] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 2

[44] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 5

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 2

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[48] R. Roussel and L. De Luca. An approach to build a complete digital report of the notre dame cathedral after the fire, using the aioli platform. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-2-2023:1359–1365, 2023. 4

[49] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11):1947, 2023. 4

[50] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[51] Farhana Sultana, Abu Sufian, and Paramartha Dutta. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowledge-Based Systems*, 201:106062, 2020. 3

[52] Guotai Wang, Jianghao Wu, Xiangde Luo, Xinglong Liu, Kang Li, and Shaoting Zhang. Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. *arXiv preprint arXiv:2306.16925*, 2023. 3

[53] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context, 2023. 3

[54] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. *arXiv preprint arXiv:2306.16741*, 2023. 3

[55] Su Yang, Miaole Hou, and Songnian Li. Three-dimensional point cloud semantic segmentation for cultural heritage: A comprehensive review. *Remote Sensing*, 15(3):548, 2023. 3

[56] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 5

[57] Shanwen Zhang and Chuanlei Zhang. Modified u-net for plant diseased leaf image segmentation. *Computers and Electronics in Agriculture*, 204:107511, 2023. 3

[58] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 5

[59] Fuzhen et al. Zhuang. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2

[60] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 5