— Supplementary Material —
Hierarchical 3D Scene Graphs Construction Outdoors

Jon Andri Nyffeler[1] Marc Pollefeys[1,3] Federico Tombari[2] Daniel Barath[1,2,4]
[1]ETH Zürich [2]Google [3]Microsoft Spatial AI Lab [4]HUN-REN SZTAKI
jon.nyffeler@gmail.com tombari@in.tum.de marc.pollefeys@inf.ethz.ch dbarath@ethz.ch

A Appendix

A.1 Example Sub-Scene Pair

Figure 1 shows an example pair of sub-scenes from the LIN dataset used for scene alignment. The overlapping region between the two sub-scenes is highlighted in orange, with each object assigned a slightly different shade. A few objects are labeled as illustrative examples. SGAligner [5] generates embeddings for each object and identifies the most similar pairs across the sub-scenes. Matched objects are connected with dotted lines for visualization.

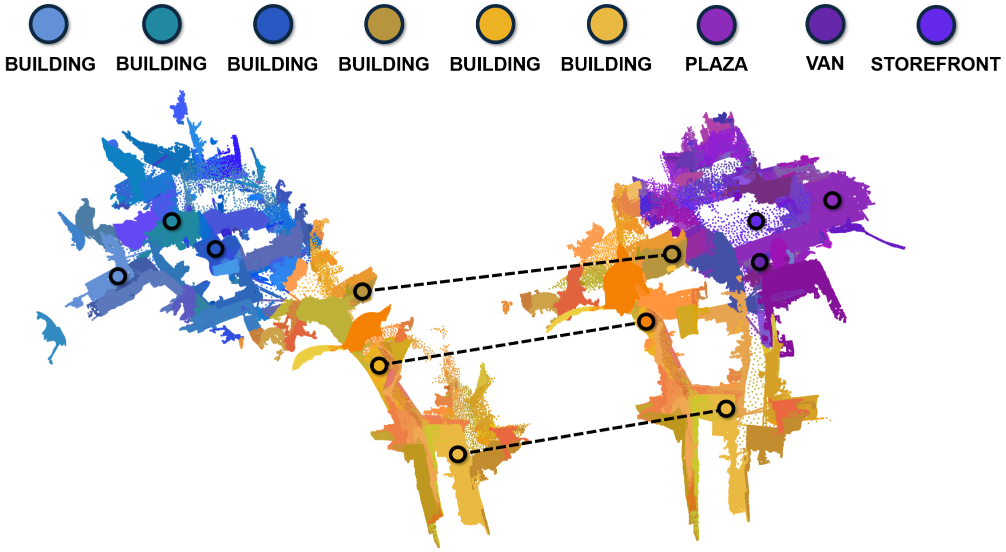
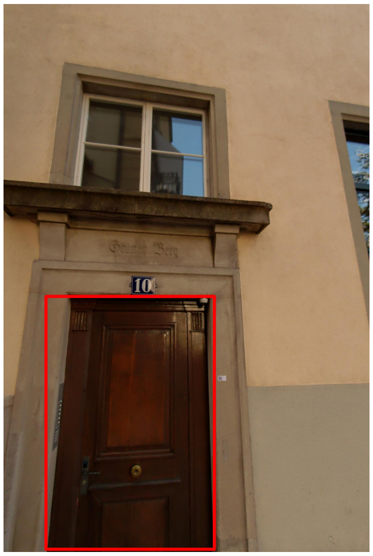


Figure 1. **Example of Object Matching for a Sub-Scene Pair.** The overlapping region is shown in orange, with distinct areas in blue and purple. Objects have slightly different tones, and selected examples are labeled with dotted lines indicating their matches.

A.2 LVLMs

A crucial step in 3D scene graph generation involves producing object labels, descriptions, and attributes. To achieve this, we select images of the target object, apply a red bounding box based on its SAM mask, and feed the result into a large vision-language model (LVLM) using a prompt designed to elicit detailed descriptions. In our experiments, we used GPT-4V [1] as the LVLM, although any other LVLM could be employed. Table 1 shows an example of object description generation using several different LVLMs. Notably, GPT-4V, Gemini [3], and Claude [2] produce similarly detailed and accurate outputs. BLIP-2 [4], the only open-source model among them, also produces reasonable results, though its descriptions and attributes are generally less precise than those of the proprietary models.



Model	Label	Description	Attributes
GPT4-V	Door	A dark wooden door with a brass peephole and metal handle, set in a stone frame.	wooden, dark, brass, handle, rectangular
BLIP2	Door	The door is brown	open, closed, window, door, windows
Gemini	Door	A brown wooden door with vertical panels and a brass doorknob is set within a stone frame below a building number.	brown, wooden, paneled, brass, doorknob
Claude	Door	A wooden entrance door with paneled design, brass hardware, and a keyhole, set within a stone doorframe.	wooden, brass, paneled, brown, entrance, solid, traditional

Table 1. **Comparison of Different LVLMs in Object Description Generation.** A door is used as the example object. On the left, the image shows the selected object with a red bounding box derived from its SAM mask. On the right, outputs from different LVLMs are shown for label, description, and attribute generation using the same prompt.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Anthropic. Claude 3 models. `https://www.anthropic.com/news/claude-3-family`, 2024. Accessed: 2025-07-31. 2

[3] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 2

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[5] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Sgaligner: 3d scene alignment with scene graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21927–21937, 2023. 1