3D Reconstruction and Semantic Labeling of Scenes from House-tour Videos

Authors: Dario Tenore, Keyshav Mor, Piyumi Jasin Pathiranage
Supervisors: Daniel Barath, Iro Armeni
ETH Zurich

Motivation

Problem Statement:

- 3D reconstruction from videos doesn't use information in audio captions directly.
- The 3D model doesn't leverage additional information embedded in audio captions.
- Utilising audio information from video can have numerous advantages in practical applications such as robotics, virtual reality, etc.

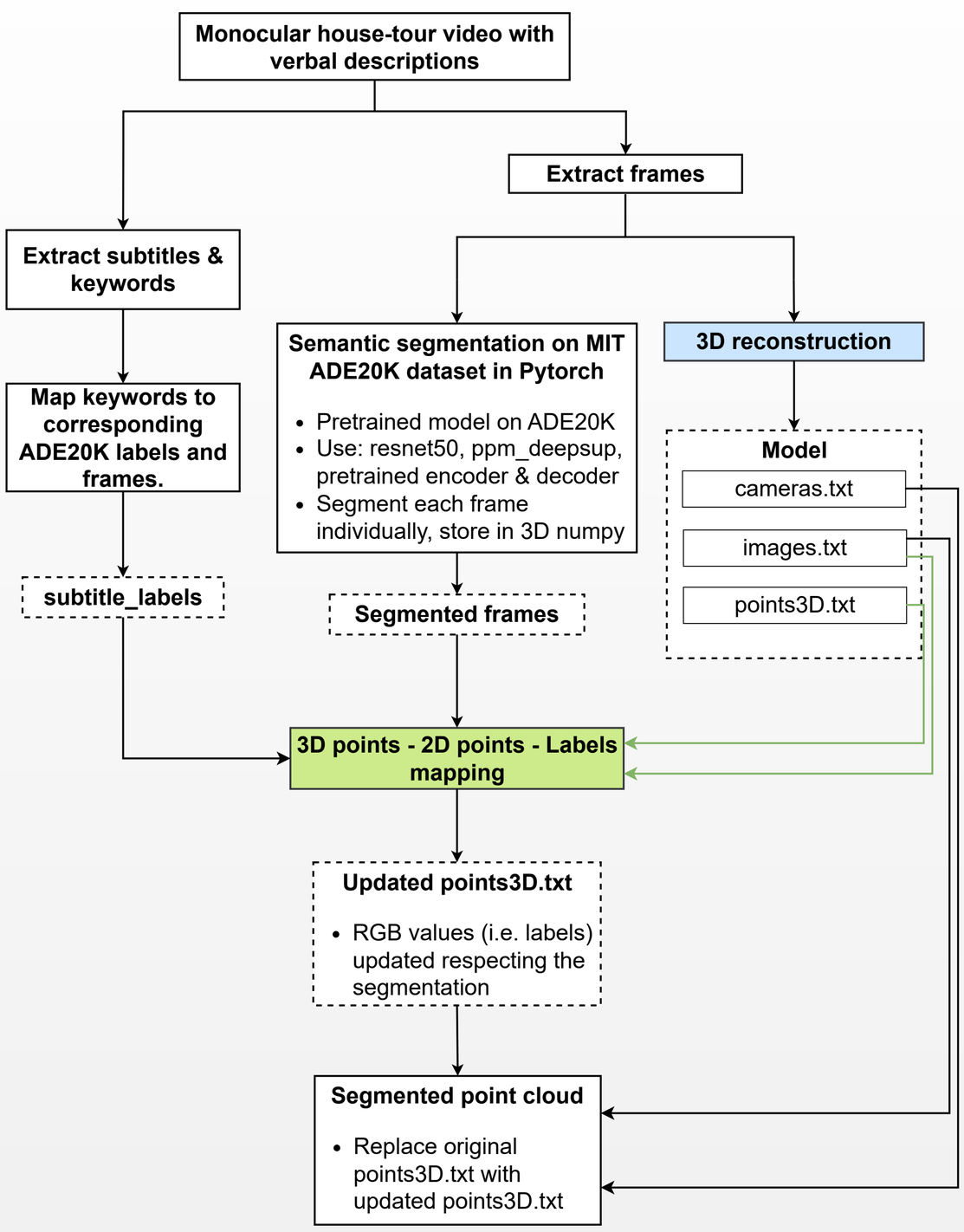
Our Proposed Solution:

- Use information from the audio captions of house-tour videos to create more semantically descriptive representations of indoor environments.

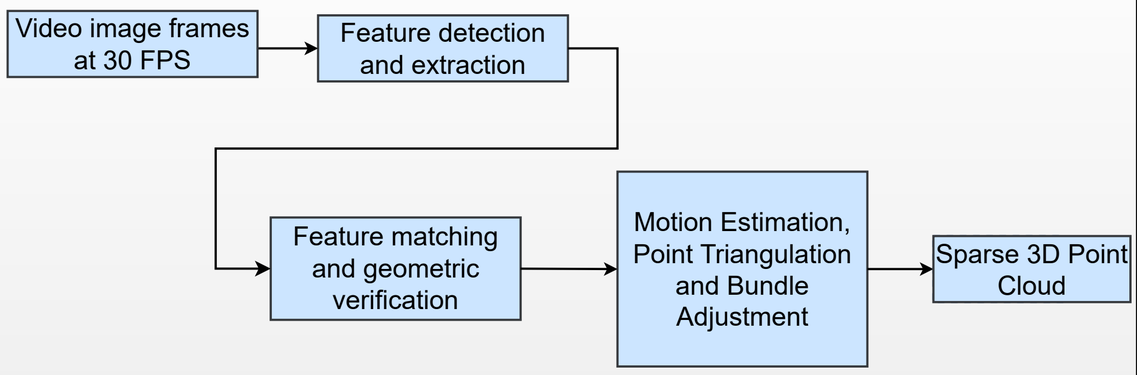
Our Contributions

- A methodology to map video subtitles to each video images frame accurately
- A methodology to extract subtitle keywords describing indoor objects and utilise them to semantically segment video image frames using CSAILVision.
- A methodology to use COLMAP 3D-2D point mapping and to backproject semantically segmented pixels back to the 3D point cloud for point-cloud segmentation.

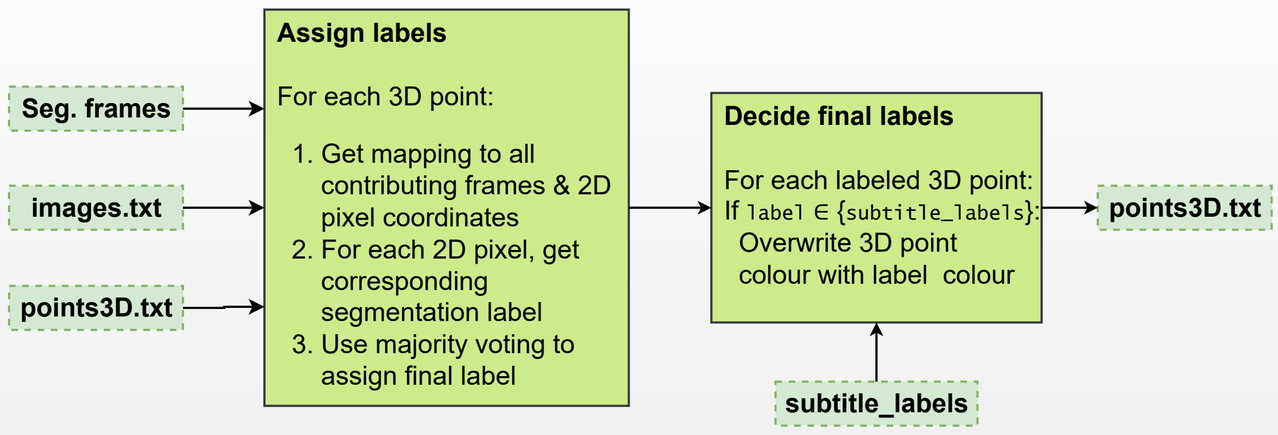
Processing Pipeline



COLMAP: 3D Reconstruction Workflow

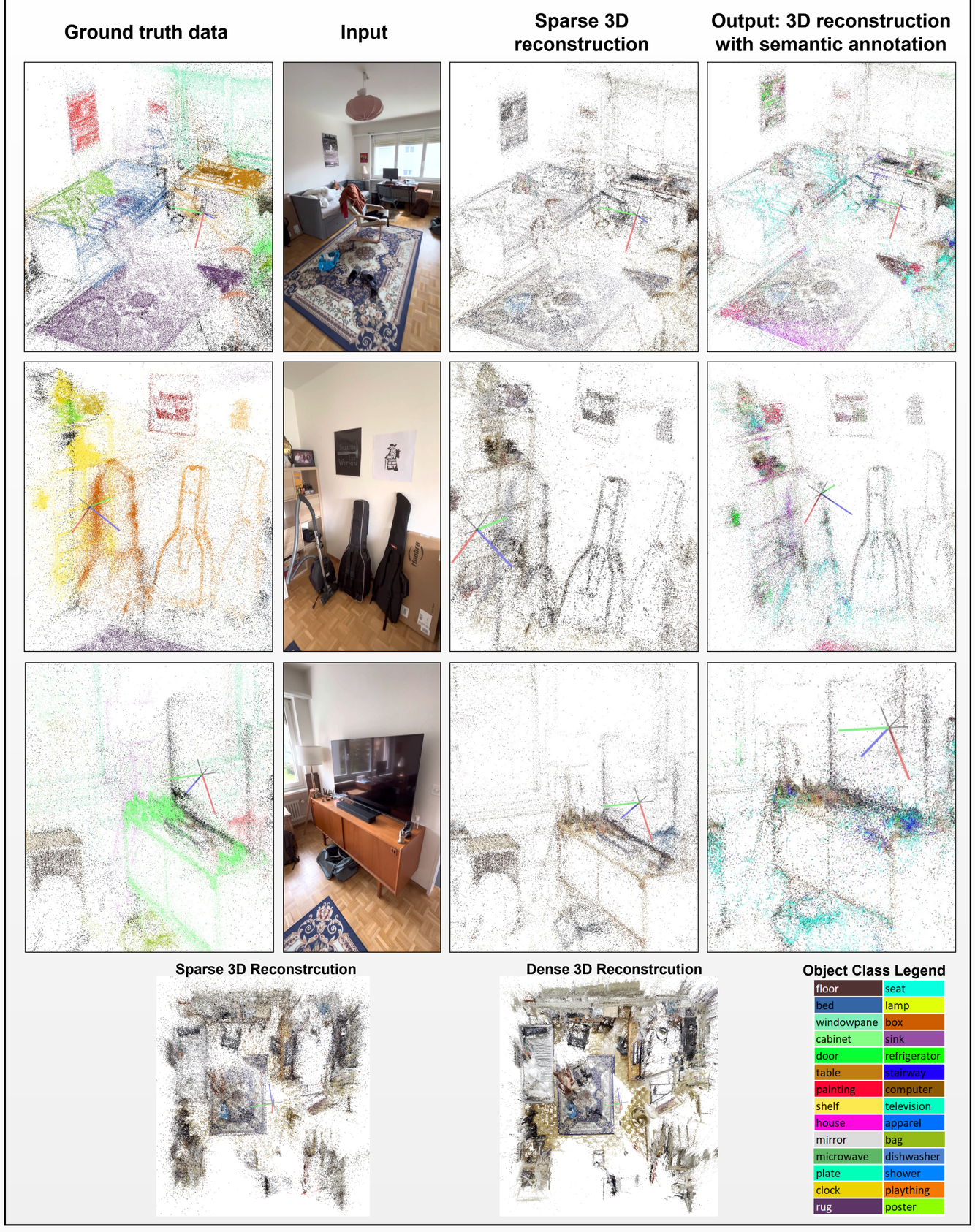


3D Points - Labels Mapping



Results

- **Qualitative results:**



Conclusion

- Our methodology enables the creation of a semantically segmented 3D point cloud of an indoor environment using verbal descriptions from a video.
- With our methodology, it is possible to annotate rich information, provided the captions include richer descriptions.
- **Limitations**: The use of a pre-existing dataset (ADE20K) limits the generalizability to environments that are not well-represented in the dataset.
- **Future work**: Improve semantic segmentation of image frames using subtitle keywords and integrate a query system to find objects using rich-descriptions from video subtitles and audio description.

References

1. COLMAP: https://colmap.github.io/index.html
2. Semantic Understanding of Scenes through ADE20K Dataset: https://arxiv.org/pdf/1608.05442.pdf
3. OpenScene: 3D Scene Understanding with Open Vocabularies: https://arxiv.org/abs/2211.15654
4. ConceptFusion: Open-set multimodal 3D Mapping: https://arxiv.org/abs/2302.07241
5. Structure-from-Motion-Revisited: https://ieeexplore.ieee.org/document/7780814,
6. Pixelwise View Selection for Unstructured Multi-View Stereo: https://demuc.de/papers/schoenberger2016mvs.pdf

ETH zürich

Computer Vision and Geometry Lab