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

Piyumi Jasin Pathiranage
ETH Zurich
ppathiran@student.ethz.ch

Keyshav Mor
ETH Zurich
keymor@student.ethz.ch

Dario Tenore
ETH Zurich
dtenore@student.ethz.ch

Abstract

This paper presents a novel approach for 3D reconstruction and semantic labeling of scenes from house-tour videos. The proposed methods leverage the textual information available from the embedded audio and subtitles of the videos to segment the 3D point clouds generated from the video. Four different frameworks are explored, leveraging the MIT CSAIL Segmentation and Open-Vocabulary Segmentation (OVSeg) for COLMAP reconstruction as well as for back-projecting RGB data to 3D points from segmented image pixels using 2D-3D correspondences. We also evaluate the Intersection over Union (IoU) of the ground-truth point clouds and the segmented point clouds to assess the accuracy of our results. Among the tested methods, the OVSeg method with back-projection of RGB data from 2D-3D correspondences results in the highest mean IoU of 0.1839. The methods we present in this paper are bootstrapped by utilising keywords from the subtitles for the segmentation and do not require expensive state-of-the-art computational setups. Furthermore, they do not rely on the additional use of CLIP to learn pixel-wise features of the images. The results demonstrate the effectiveness of our proposed approaches in achieving context-aware semantic labeling on a simple computing setup, with potential applications in scene understanding, robotics, and virtual reality.

1. Introduction

Accurately understanding and representing the 3D structure of real-world scenes is a fundamental task in computer vision and robotics. Semantic 3D reconstruction plays a crucial role in various applications such as scene understanding, robot navigation, and virtual reality.

In traditional approaches, semantic segmentation of 3D reconstructions heavily relies on visual information, often overlooking the additional sources of information available in the multimedia data. House-tour videos, for instance, provide an additional wealth of data in the form of audio captions, which offer rich descriptions and contextual cues about the objects within the scene.

The motivation behind this work stems from the recognition that by incorporating these underutilized audio captions, we can significantly enhance the segmentation of point clouds and improve the overall understanding of scenes. By leveraging the textual information, including object descriptions, and contextual hints, we aim to segment the point cloud in a manner that utilises the additional information of the objects that we have. Moreover, with the incorporation of audio and caption information, we aim to enhance the robustness of the semantic segmentation approach such that it can achieve more reliable and accurate segmentation results when combined with visual information used for 3D reconstruction.

The potential benefits of this approach are vast. It allows for more precise and context-aware semantic labeling such as enabling robotic systems and virtual reality applications to locate, navigate and interact with specific objects in the real world using a 3D reconstructed scene map.

Our study makes several noteworthy contributions:

- **Leveraging subtitles for improved segmentation:** Our work focuses on utilizing keywords in the video subtitles and the keyword-to-frame mapping as a valuable information source for segmenting the 3D scene.

- **Technology-efficient semantic segmentation methods:** We introduce four methods that enable the semantic segmentation of 3D scenes that do not need state-of-the-art expensive computational infrastructure and that utilise zero-shot segmentation of video frames using MIT CSAIL Segmentation [10, 9] and Open-Vocabulary Segmentation (OVSeg) [3]. These methods provide practical alternatives for users with limited hardware resources, allowing them to perform semantic segmentation tasks using a simplified setup.

- **Back-projection of segmented RGB pixel data from 2D to 3D:** We present a methodology to back-project segmented RGB pixel data from 2D images to 3D points using the 2D-3D point correspondences generated by COLMAP.

2. Related Work

In the field of 3D computer vision, extensive research has been conducted in the semantic segmentation of 3D scenes and 3D scene understanding.

Songyou Peng et al. [5] propose OpenScene which predicts dense features for 3D points that are co-embedded with text and image pixels in CLIP [6] feature space. The main distinction between our study and OpenScene is that we leverage textual descriptions in subtitles to enhance the segmentation , whereas they utilise the CLIP [6] features.

Another closely related work is ConceptFusion [2] which utilizes RGB images, depth data, and CLIP [6] or DINO [1] features to create an open-set multi-modal 3D scene. They demonstrate the fusion of pixel-aligned open-set features into 3D maps using traditional SLAM and multi-view fusion approaches. Unlike ConceptFusion, our approach uses textual keyword descriptions from video subtitles to semantically segment 2D images in a zero-shot approach using MIT CSAIL Segmentation and OVSeg.

Unlike these works, our approach is inherently closed-set and cannot be used for open-set queries.

3. Methodology

In this section, we present the four methodologies employed in our study. We provide a brief overview of these methodologies, highlighting their differences, followed by detailed descriptions of each combination.

3.1. Overview of Method Combinations

- **MIT CSAIL + COLMAP**: This method involves utilizing the MIT CSAIL semantic segmentation method [9, 10] to segment the frames of the house-tour video using the keywords extracted from the subtitles, followed by 3D reconstruction of the segmented frames using COLMAP [7, 8].
- **MIT CSAIL + Back-projection**: In this method the MIT CSAIL semantic segmentation method [9, 10] is employed to segment and mask the frames. The masked frames are then flattened into arrays that get used for the back-projection (Figure 1) of 2D RGB data, where the labels and their corresponding RGB values are assigned to the corresponding 3D points in an already reconstructed 3D point cloud.
- **OVSeg + COLMAP**: This method utilizes the Open-Vocabulary semantic segmentation [3] approach to segment the frames of the video using keywords extracted from the subtitles, followed by 3D reconstruction of the segmented frames using COLMAP [7, 8].
- **OVSeg + Back-projection**: This combination uses Open Vocabulary segmentation [3] for keyword-based segmentation and masking of frames. The masked frames are then flattened into arrays that get used for the back-projection (Figure 1) of 2D RGB data, where the labels and their corresponding RGB values are assigned to the corresponding 3D points in an already reconstructed 3D point cloud.

3.2. Keyword Extraction and Data Processing

The data processing step involves extracting subtitles from the audio of the house-tour videos, establishing a frame-to-subtitle mapping, and performing keyword extraction using a natural language processing, the NLTK framework. We focus on extracting nouns as they typically represent objects in the scenes. These keywords serve as important semantic cues for subsequent segmentation and labeling processes.

Both the segmentation frameworks we use i.e. MIT CSAIL Segmentation and OVSeg are evaluated on the ADE20K dataset [9, 10], thus we decided to use the same dataset. The ADE20K dataset contains a large variety of indoor and outdoor scenes, with around 150 object labels. However, to improve segmentation accuracy in the house-tour context, we map around 40 object labels from the ADE20K dataset to a subset of 11 closely related final labels to make the segmentation more robust. For each extracted keyword, we assign the most appropriate final label based on a predefined mapping table (refer to Figure 3h) where we also assign each label a colour that is to be used for the ground-truth definition and segmentation.

3.3. MIT CSAIL + COLMAP

This method combines the pixel-wise object labeling capability of MIT CSAIL Semantic segmentation with the robust 3D reconstruction capabilities of COLMAP. The MIT CSAIL semantic segmentation is trained on the ADE20K using multiple encoder-decoder deep learning models to perform pixel-wise segmentation of objects in each frame. This process involves feeding the frames into the MIT CSAIL segmentation model, utilizing the selected set of 11 final labels for segmentation, and suppressing all other labels by setting their masks to an RGB value of [0 0 0].

Once the segmentation is obtained, the segmented masks are overlaid on the original frames with the transparency for the alpha channel of the image set at 0.5. This overlay visually combines the segmented masks of the objects with the original scene, preserving some features of the original image useful for 3D reconstruction. An error analysis of this approach is provided in a later section to evaluate how many 3D points are correctly segmented using this approach.

The next step involves reconstructing the 3D point cloud using COLMAP from the overlaid segmented frames. COLMAP utilises SIFT [4] feature extraction and matching, and point triangulation for reconstructing a 3D scene using images of a camera in motion. In this process, the overlaid segmented frames serve as input to COLMAP which generates a sparse point cloud representation. The

reconstructed 3D scene contains points that contain a hue that is close to the hue of the masking RGB label used in the video frame segmentation. This hue helps us compute the Intersection over Union (IoU) with the ground truth point cloud. The reconstructed 3D scene is manually segmented using predefined RGB object labels to generate the ground truth which is eventually used to compute the IoU with the semantically segmented 3D scene.

3.4. MIT CSAIL + Back-projection

This method involves applying the MIT CSAIL segmentation model to each frame using the chosen 11 labels. Instead of using the resulting segmented frames directly for 3D reconstruction in COLMAP, we utilize the original unmasked frames (i.e., prior segmentation) for the 3D reconstruction using COLMAP. We also manually segment this reconstruction with the pre-defined RGB labels to define the ground-truth point cloud. Finally, the back-projection technique is employed to assign RGB labels to the original reconstructed 3D point cloud. The next section will delve into the intricate specifics of the back-projection technique.

3.4.1 Back-projection Technique

To enhance the semantic labeling of the reconstructed 3D point cloud, a back-projection algorithm is employed. The algorithm associates each 3D point with an RGB label based on the majority voting of the semantically segmented 2D points, choosing the most common label associated with the 2D point and projecting its RGB label onto the corresponding 3D point. Figure 1 shows a summary of this process.

3.5. OVSeg + COLMAP

The Open Vocabulary segmentation method [3] provides an alternative approach to semantic segmentation. It takes open-vocabulary inputs like keywords from video subtitles to generate CLIP feature embeddings and then looks for objects in the image with similar CLIP embeddings. If it finds a match, it assigns the area of interest the highest score and semantically segments these objects with a predefined RGB mask. We modify this process a bit by making the masks transparent by setting the alpha channel to 0.5 and by defining the RGB values for the masks associated with our keyword label queries. The transparent mask helps us in preserving features of the original image useful for 3D reconstruction. Following the segmentation step, the segmented frames are fed into COLMAP to perform the 3D reconstruction. The 3D reconstruction results in the semantically segmented 3D point cloud. This point cloud is also manually segmented with predefined RGB label masks to ensure an accurate IoU which can be obtained vis-a-vis the semantically segmented point cloud.

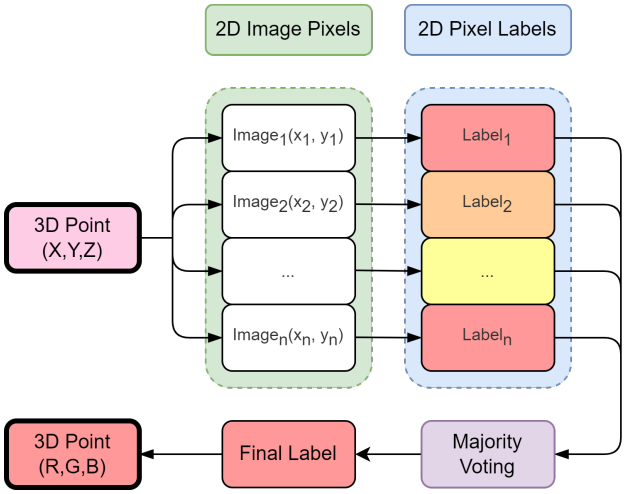


Figure 1: For each 3D point, we go over all contributing 2D pixels of all corresponding images. When then gather the labels of these corresponding pixels only if their label class was mentioned in the subtitles. Next, we do majority voting on those labels. As a last step, we change the original 3D point's RGB value to the one corresponding to the final class label, as can be seen in the Legend 3i.

3.6. OVSeg + Back-projection

This method involves applying the OVSeg segmentation to each frame using the chosen 11 final labels. Instead of using the resulting segmented frames directly for 3D reconstruction in COLMAP, we utilize the original unmasked frames for the 3D reconstruction using COLMAP. We also manually segment this reconstruction with the pre-defined RGB labels to define the ground-truth point cloud. Finally, the back-projection technique is employed to assign RGB labels to the original reconstructed 3D point cloud.

3.7. Hardware Setup and Configuration

For running our computation pipelines, we used commercial off-the-shelf GPUs and CPUs which are common for domestic use. Our machine configurations used an NVIDIA RTX3080 12GB, AMD 5900X with 32GB of RAM and an NVIDIA RTX 3070 8GB, AMD Ryzen 7 3700X with 32 GB of RAM respectively. Both setups could be used alternatively for our methods.

4. Results

Our dataset comprised a self-recorded house-tour video, accompanied by verbal descriptions of the various objects present in the house. To evaluate the performance of the examined methods, we conducted four experimental runs, applying each of our for methods. We computed the IoU of all the segmented 3D point clouds with their respective ground

truth point clouds to assess the efficacy of each method's segmentation outcomes.

To gain initial insights into the performance, we first present the outcomes of the two different segmentation techniques, MIT CSAIL and OVSeg (refer to Figure 2). Upon visual examination, it is evident that the segmentation results of the MIT CSAIL algorithm (Figure 2b) are significantly poor. This eventually reflects in the segmentation and IoU of the MIT CSAIL Segmentation with COLMAP and Back-projection. Conversely, the OVSeg segmentation algorithm (Figure 2d) demonstrates superior performance, exhibiting highly precise and detailed segmentation outcomes for the shelf, the lamp, the clock and the paintings.

Thereafter, we present visual representations of the segmented point clouds generated by each method with their corresponding ground truth data (refer to Figure 3).

4.1. Evaluation Metrics

To compare the accuracy of the methods, we computed Intersection over Union (IoU) scores for each label and the mean IoU across all labels. Two distinct approaches were employed to compute the IoU, tailored to the methods with back-projection and the methods with COLMAP reconstruction, respectively. The inclusion of two different approaches for computing IoU scores stems from observations made during the 3D reconstruction phase using COLMAP. In the COLMAP-generated 3D point cloud of semantically segmented video frames, slight variations in RGB intensity were noted when compared to the ground truth labels. This can be attributed to the fact that COLMAP integrates information from multiple image pixels to generate the final 3D points. Consequently, the RGB values of the reconstructed points do not precisely match the RGB values of the ground truth labels. In contrast, the back-projection technique directly alters the RGB values of the pixels, resulting in a closer correspondence between the assigned labels and the true labels. To account for these differences in RGB intensity and accurately evaluate the segmentation performance, separate approaches were employed to compute the IoU scores for the methods using COLMAP [7, 8] and back-projection. Here, we provide a detailed explanation of each approach.

For the methods involving back-projection, we scan all the 3D points and if the RGB value of the 3D points matches with any of the ground-truth RGB labels, we assign them the corresponding label, such as "Bed" and "Chair". We then computed individual IoU scores for each labelled object with the ground truth data, by finding the closest points within a proximity threshold of 0.0001 and then selecting points with cosine similarity of above 0.9 as intersecting points. Based on the intersection points, we were able to yield the IoU values.

For the methods involving COLMAP reconstruction of segmented images, a comparison was made between each RGB value of every 3D point in the point cloud and the set of 12 RGB values corresponding to the labels. The cosine similarity was computed for each comparison, and the label associated with the highest cosine similarity was assigned to the respective RGB value. Subsequently, the same proximity checking and cosine similarity thresholding steps as in the previous approach were employed to compute the IoU scores.

The computed IoU scores for each label and the mIoU for the different methods are presented in Table 1.

Table 1: IoU values of the segmentation results for each label and the mean IoU (CSAIL1: CSAIL+ Backprojection; OVSeg 1: OVSeg + COLMAP; CSAIL 2: CSAIL + COLMAP; OVSeg 2: OVSeg + Backprojection

Label Class	CSAIL 1	OVSeg 1	CSAIL 2	OVSeg 2
Bed	1.01%	0.82%	0.02%	**12.73%**
Shelf	0%	10.58	3.16%	**26.08%**
Table	1.67%	1.16%	0.09%	**8.04%**
Chair	0.93%	1.12%	0.16%	**18.02%**
Cabinet	9.13%	2.46%	0.65%	**15.14%**
Lamp	3.48%	11.17%	0.2%	**25.37%**
Screen	0.58%	**14.96%**	0.21%	14.42%
Cloth	0.85%	2.77%	1.12%	**11.41%**
Painting	7.15%	0.76%	0.64%	**17.05%**
Clock	8.65%	3.81%	0.52%	**32.9%**
Rug	1.09%	35.31%	**40.01%**	21.08%
mIoU	3.14%	7.72%	4.25%	**18.39%**

4.2. Comparative Analysis

Analyzing the results presented in Table 1, we can derive several interesting insights regarding the segmentation performance of the four different methods of the 11 labels.

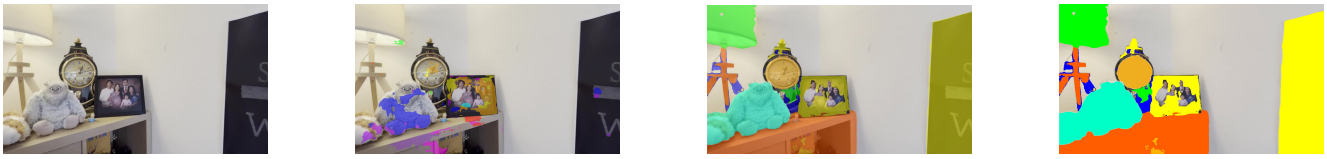
4.2.1 Method Performance

The combination of Open Vocabulary segmentation with our back-projection technique yielded the highest mean IoU of 0.1839, indicating a relatively better overall performance compared to the other combinations.

When assessing the label-specific IoU values, OVSeg + Back-projection achieved the highest scores for 9 out of 11 labels, exhibiting only a marginal underperformance for the label Screen and a significant underperformance vis-a-vis MIT CSAIL COLMAP for the label Rug. OVSeg with COLMAP reconstruction shows us that the segmentation can be improved in combination with COLMAP if the semantic segmentation of input images is accurate enough.

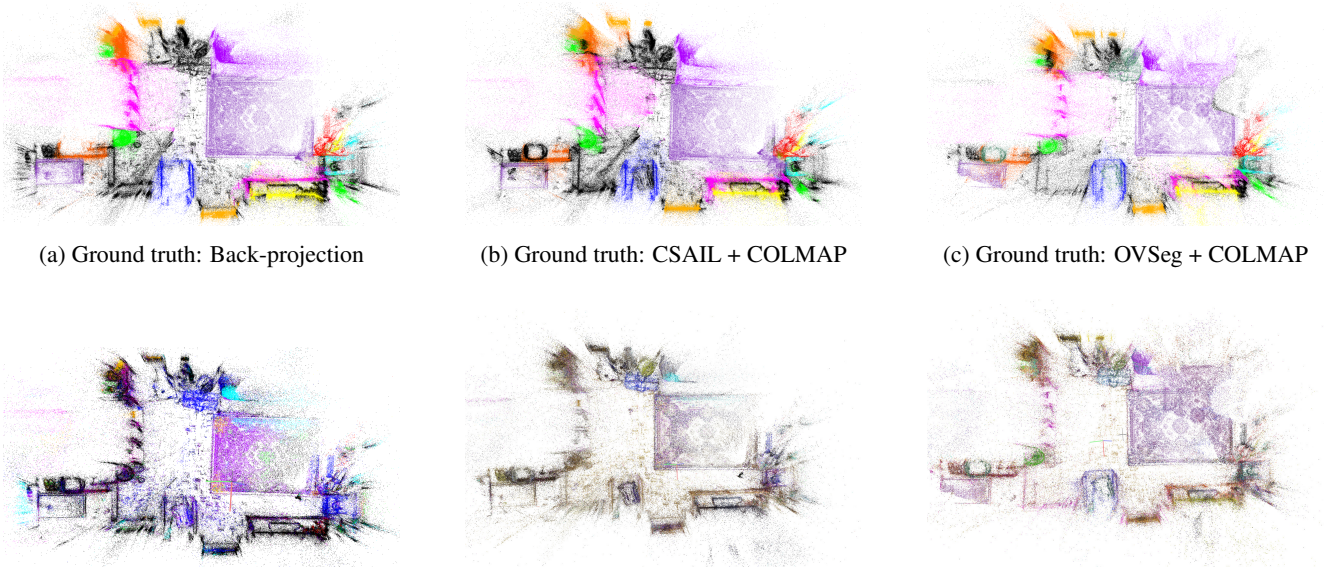
5. Discussion

Looking at Table 1, it is evident that the combination of OVSeg + Back-projection consistently yields higher IoU



(a) Original image (b) CSAIL segmentation (c) OVSeg transparent (d) OVSeg opaque

Figure 2: An example of the segmented frames.



(a) Ground truth: Back-projection (b) Ground truth: CSAIL + COLMAP (c) Ground truth: OVSeg + COLMAP

(d) CSAIL + Back-projection (e) CSAIL + COLMAP (f) OVSeg + COLMAP

(g) OVSeg + Back-projection

Keyword	Label	Keyword	Label	Keyword	Label
curtains	cloth	light	lamp	rucksack	bag
blanket	cloth	lamps	lamp	sideboard	cabinet
cushion	bed	clock	clock	cupboard	cabinet
mattress	bed	seat	chair	tv	screen
pillow	bed	armchair	chair	monitor	screen
shelf	shelf	chair	chair	screen	screen
books	shelf	leg	table	computer	screen
carpet	rug	poster	painting		

(h) Keyword to final label mapping.

(i) Legend

Figure 3: Segmented point cloud of different combinations with the corresponding ground truth.

values compared to the other methods for most label classes. This suggests that the integration of Open Vocabulary segmentation with the back-projection technique leads to more accurate and robust object segmentation and labelling.

One possible reason for the superior performance of OVSeg + Back-projection is the effectiveness of Open Vocabulary segmentation in handling open-vocabulary keyword queries. By utilizing an open vocabulary approach, the system can adapt to new or previously unseen objects during the labelling process. This flexibility allows for more accurate and comprehensive segmentation, resulting in higher IoU values.

The house-tour video was shot in an apartment with white walls and light-coloured objects, many of which lacked rich textures, resulting in poorer reconstruction of the objects. This also affects the identification or results in

misidentification during semantic segmentation of the video frames. This also affects the overall IoU.

We believe our method would perform well in texture-rich 3D environments where the semantic segmentation of video frames, as well as COLMAP reconstruction, can really be leveraged to their full potential.

5.1. Factors Influencing IoU Values

Examining the reasons behind the lower IoU values for other combinations, several factors come into play. One factor is the potential confusion of objects during the segmentation process. For instance, OVSeg [3] may mistakenly classify doors as cabinets or cupboards, leading to discrepancies between the segmented output and the ground truth labels, resulting in lower IoU values.

Additionally, the usage of multiple terminologies for the same object can contribute to reduced IoU values. For example, clothes can be classified as "Clothes" (the commonly used term) or as "Apparel" (the label in the ADE20K dataset). Similarly, different levels of detailing can influence the segmentation results. For instance, trousers can be classified as "Clothes" or "Trousers" based on the desired level of segmentation detail. These factors contribute to the variations in IoU values observed across different labels. The similarity in the RGB mask of "painting" and "clock" might have also affected their IoU values.

In the keyword-to-final-label mapping (refer to Figure 3h) we tried to mitigate these characteristics of the ADE20K labels and assigned the most suitable label based on our judgment. However, it is important to acknowledge the potential presence of bias in this assignment process.

5.2. Evaluation Metrics Analysis

In Table 2 we present an analysis of the performance of the transparent RGB masks on video frames in segmenting the 3D point cloud when the frames are fed to COLMAP.

The "Correctly Seg." columns represent the ratio of points having a cosine similarity above 0.9 (when compared with the corresponding RGB ground-truth label) and the proximate points within a threshold of 0.0001 (compared to the ground truth points). The "Correctly Intersec." columns represent the ratio of points that actually intersect (between the ground truth and the segmented point cloud) and the number of manually segmented ground-truth points.

As visible from the mean values, the OVSeg + COLMAP reconstruction achieves almost 3 times the performance of CSAIL+COLMAP reconstruction when segmenting and achieves almost 50% better performance in achieving intersection with the ground truth.

This observation supports our theory that our method would perform well in texture-rich 3D environments where the semantic segmentation of video frames, and COLMAP reconstruction, can be leveraged to their full potential.

Table 2: Percentage of points correctly segmented and intersecting with ground truth when using transparent RGB masks on segmented images for COLMAP (Left two columns: OVSeg; Right two columns: CSAIL)

Label	Segment	Intersect	Segment	Intersect
Bed	15.35%	0.85%	1.05%	0.02%
Cabinet	14.18%	2.9%	6.4%	0.73%
Chair	4.41%	1.48%	0.27%	0.4%
Clock	3.9%	62.4%	1.25%	57.74%
Cloth	15.76%	3.25%	1.35%	6.44%
Lamp	65.75%	11.87%	3.97%	0.21%
Painting	30.15%	7.79%	5.97%	0.71%
Rug	43.3%	65.66%	60.12%	54.48%
Screen	50.99%	17.48%	1.23%	0.25%
Shelf	32.59%	13.56%	10.95%	4.26%
Mean	**25.81%**	**16.50%**	**08.35%**	**11.39%**

6. Future Work

In future work, we would like to test our methods on environments with rich chromatic and textural information to evaluate the performance of our methods. We would also like to evaluate our methods for dense 3D reconstructions. Another aspect to consider for future work is the implementation of a query system to query objects and rich descriptions, ensuring that they accurately represent the intended labels. We would also like to evaluate our methods to create a multi-level 3D scene graph of a multi-storeyed 3D environment and see the performance of our methods.

7. Conclusion

In this study, we have presented a comprehensive methodology for scene understanding and 3D reconstruction of house-tour videos. By combining MIT CSAIL semantic segmentation [9, 10], Open Vocabulary segmentation [3], COLMAP reconstruction [7, 8], and back-projection techniques, we have achieved good segmentation of 3D scenes using a domestic computing infrastructure and a simple pipeline that utilises keywords from video subtitles.

Through our experiments on a house-tour video, we have evaluated the performance of different methods using IoU metrics and error analysis of segmented and intersecting points when using transparent masks on segmented images. Our findings provide valuable insights into the strengths and limitations of different methods. Our methods also prepare the groundwork for integrating audio embeddings of videos to generate robust semantically segmented 3D scenes.

8. Work Distribution

Dario Tenore created the pipeline for the MIT CSAIL segmentation and the back-projection pipeline used for two of the methods. Keyshav Mor did all the reconstructions in COLMAP, set up the OVSeg pipeline to generate the transparent and opaque segmentation masks for video frames and wrote the code for the IoU calculations and segmentation error analysis. Piyumi Pathiranage created the subtitle-keyword extraction pipeline and the corresponding label mappings.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[2] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping, 2023. 2

[3] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip, 2023. 1, 2, 3, 6

[4] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–, 11 2004. 2

[5] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies, 2023. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[7] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 6

[8] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 4, 6

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 6

[10] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 1, 2, 6