國立清華大學 電機工程學系
實作專題研究成果摘要

CLIP-Driven Region Growing for End-to-End 3D Semantic Instance Segmentation
基於 CLIP 引導的區域成長方法用於端到端 3D 語義實例分割

專題領域：資工領域

組別：A532

指導教授：孫民

組員姓名：周冠霖

研究期間：113 年1 月 1 日至 114 年 5 月 1 日止 共 15 個月

Abstract

With the advancement of 3D instance segmentation technologies, many studies have relied on annotated datasets for training. OV3DIS (Open Vocabulary 3D Instance Segmentation) aims at resolving open-set semantic reasoning. SAI3D has made great progress by combining geometry-based primitives with 2D indexed mask labels from SAM. SAI3D first generates 3D class-agnostic masks by producing primitives, projecting them onto 2D, using Semantic-SAM to generate 2D indexed mask labels (e.g., 0, 1, 2). Then compute affinity scores based on indexed mask labels and then apply a region-growing approach to obtain class-agnostic 3D instance masks. To achieve open-vocabulary segmentation, SAI3D adopts the OpenMask3D pipeline: the 3D class-agnostic masks are first projected back to 2D, then processed with cropping, SAM, and CLIP, and finally aggregated into a 3D instance-level CLIP feature for semantic assignment.

However, it is evident that the repeated projections and the two-time use of SAM introduce significant computational overhead, prompting me to explore whether a more streamlined pipeline could accomplish the same task. My goal is to improve the end-to-e d process and generate better 3D mask CLIP features.

To address these issues, I propose a new method that integrates CLIP features at an earlier stage, allowing the generated 3D instance masks to carry semantic information, and utilizing the CLIP feature for better region growing process. These innovations can enhance the accuracy of class agnostic segmentation and better end-to-end semantic segmentation pipeline, showing strong potential for improved 3D scene understanding.

摘要

隨著 3D 實例分割(Instance Segmentation) 的發展，許多研究依賴標註數據集進行訓練。OV3DIS（Open Vocabulary 3D Instance Segmentation）旨在解決開放集合語義推理(open-set semantic reasoning) ，常見的做法是利用視覺語言模型來進行語義推理，突破了傳統方法在物體類別上的限制。SAI3D 通過結合幾何基元(geometric primitive) 和來自 SAM 的2D 索引式遮罩標籤(indexed mask labels)，在此領域取得巨大進步。

SAI3D 原本的流程是生成幾何基元(geometric primitive)、將其投影到 2D、使用 semantic-SAM 生成 2D 索引式遮罩標籤(indexed mask labels) (例: 0, 1, 2)，並以 2D 索引式遮罩標籤(indexed mask labels) 計算相似度分數，接著應用區域生長(region growing) 生成 3D 無類別區分實例遮罩(class agnostic instance masks)。而語義分配則是交由 OpenMask3D，將 3D 無類別區分實例遮罩(class agnostic instance masks) 投影至 2D，經由裁切、SAM、CLIP，將不同視角整合成 3D 實例遮罩 CLIP 特徵(instance mask CLIP feature)。

但顯而易見的，此種做法將會使用兩次的投影以及 SAM，故我提出更精簡的流程來完成這個任務，期望能改善計算成本，並嘗試產生出更好的 3D 實例遮罩 CLIP 特徵(instance mask CLIP feature)。

針對上述問題，我提出了以下改進方法: 更早的整合 CLIP 特徵，不只使得生成的 3D 遮罩已經包含語義信息，也利用 CLIP 特徵產生更好的 3D 實例遮罩。這些創新可以提升類別無關分割(class-agnostic segmentation)的準確性，並優化端對端語義分割流程，展現出增進 3D 場景理解的潛力。

章節目錄

1. Introduction 1

1.1. 研究目的 1

1.2. 背景簡介 1

1.3. 問題說明 1

1.4. 解決方法 2

1.5. 文獻探討 2

2. Research Methodology 3

2.2. Scene Graph Construction 4

2.3. Primitive Affinity 5

2.4. Primitive Merging 6

2.5. Feature aggregation. 7

3. Experimental Results 8

3.1. Experiment Setup 8

3.2. Results 9

3.3. Ablation and Analysis 10

4. Conclusion 10

5. 心得感想 11

6. Reference 11

圖目錄

Figure 1. System Design 4

Figure 2. Under-segmentation example 10

表目錄

Table 1. Class-agnostic 3D instance segmentation on ScanNetV2 dataset. 9

Table 2. Semantic instance segmentation on ScanNet200 dataset. 10

1. Introduction

1.1. 研究目的

於 3D Open Vocabulary Instance Segmentation 提出新的流程，產出一個在性能表現上具競爭力的端對端流程。

1.2. 背景簡介

在 3D Instance Segmentation 中， Open Vocabulary Instance Segmentation 越來越受到關注，與 Closed Vocabulary Instance Segmentation 不同，注重方法對訓練過程中未見過類別的泛化，而不是侷限在固定的類別清單中。常見的方法是先生成3D class agnostic masks，然後結合 2D 投影與語義模型（如 SAM[5] 和 CLIP[4]）來進行語義推理，其中 3D class agnostic masks 的品質對最終結果影響重大。

SAI3D[1] 在 3D class agnostic masks 提出了新的方法，通過幾何性質將 3D point cloud 轉換為 geometric primitives，並利用 Semantic-SAM[6] 生成 2D indexed masks ，並以2D indexed masks 計算 primitives 之間的 affinity score，進行合併以產生 3D class-agnostic instance masks。儘管 SAI3D[1] 生成的 3D masks 品質優秀，但在語義分配上仍依賴於 OpenMask3D[2]的投影與語義推理的 pipeline，這使得計算效率存在優化空間。

我的研究動機是基於這一發現，認為如果在 primitives 階段就能夠保留足夠的語義信息，那麼在合併 primitives 時便能同時生成語義豐富的3D instance feature，從而簡化整體流程並提高性能表現。

1.3. 問題說明

在給定3D point cloud 與對應的 2D posed RGB-D 影像下，我們的目標是建立一個有效的 3D Open-Vocabulary Instance Segmentation 流程，具體挑戰如下：

(1) 高效生成 point-wise CLIP 特徵：尋找兼顧效率與準確性的方式，為每個點生成語義豐富的 CLIP 特徵。

(2) 設計 primitive 間的合併邏輯，判斷是否將複數 primitives 合併為同一 instance。

(3) 確保在 primitives 合併為 3D instance masks 時，能同時獲得語義豐富的 3D instance-level CLIP 特徵。

1.4. 解決方法

(1) 參考 Grounded SAM[8] 流程，結合 GroundingDINO[9]、SAM[5] 和 RAM[10]，對每張影像產出高品質的 2D segmentation masks。接著，使用 CLIP 提取每個 2D mask 的特徵，並透過 3D-to-2D 投影將這些特徵對應到 3D point cloud 上，進而產出 point-wise 的 CLIP 特徵。

(2) 為判斷是否將兩個 primitives 合併成一個 instance，設計三項條件作為基準，僅當三者皆超過閾值才進行合併。三種條件分別為將3D primitive 投影到2D 圖片上所覆蓋區域對應到的:

i. Image CLIP feature.

裁剪影像經過 CLIP image encoder。

ii. Text CLIP feature.

GroundingDINO[9]給出， 例: “chair”，再經過 CLIP text encoder。

iii. Indexed mask labels:

SAM[5]給出的2D mask indices，例: 0, 1, 2

(3) 完成 merging 產生最後的3D 實例後，將 instance 中每個點的 CLIP 特徵做平均得到3D instance-level CLIP 特徵。

1.5. 文獻探討

(1) SAI3D[1]:

SAI3D[1]提出了一個方法，利用幾何性質先將點處理成 primitives，而後將 primitives 投影至2D，使用 Semantic-SAM[6]生成2D indexed mask labels，並以2D indexed mask labels 計算每個 primitives 間的 affinity score，最後利用計算好的

affinity score 來決定是否將複數個 primitives merge 成一個 instance，merge 過程結束後即得到 3D class-agnostic instance masks。但因為後續需要依賴 OpenMask3D[2]的 semantic assignment pipeline，兩個步驟的脫節導致執行了重複的投影以及 SAM[5]。

(2) OpenMask3D[2]:

OpenMask3D[2]的架構具有很高的泛用性，其主要依賴如 Mask3D[6]的 pretrained backbone model 來生成3D class agnostic masks，將3D class agnostic masks 投影至 2D images，經過裁切、SAM[5]、CLIP[4]得到每個2D instance mask 的 CLIP feature，最後將其整合成3D instance CLIP feature。

OpenMask3D[2]明顯的缺點就是高度仰賴3D class agnostic masks，但其分配語義的架構充分利用了收益於2D 大量資料的 CLIP[4]、SAM[5]，將其優勢轉移至3D 的任務裡，很具參考價值。

(3) Grounded SAM[8]

Grounded SAM[8] 是一種結合 Grounding DINO[9] 和 Segment Anything Model (SAM) [5] 的多功能視覺系統，具備根據任意文字提示進行物件偵測與分割的能力。這使其在開放詞彙的場景下展現了極高的靈活性與泛化能力。

其最大特色之一，是可整合不同的視覺模型以實現複雜任務。原論文中即提到，透過加入 Recognize Anything Model (RAM)[10]，可建構基於輸入影像的自動標註流程。本專題應用此標註策略生成所需的 2D indexed mask labels (例: 0, 1, 2) 以及文字語義標籤(例: “chair”)。

2. Research Methodology

System overview

Figure 1. 展現了我們方法的流程。首先將 3D point cloud 集合成 primitives，並利用 Grounded SAM[8]自動標註流程配合 CLIP[4]，生成三種所需 2D 特徵。接著利用投影產生 primitive-level 特徵並計算特徵間的相似度。以相似度為基準執行

region growing，完成合併後即可生成 class agnostic instance segmentation。再根據此結果聚合語義特徵以進行 semantic instance segmentation。

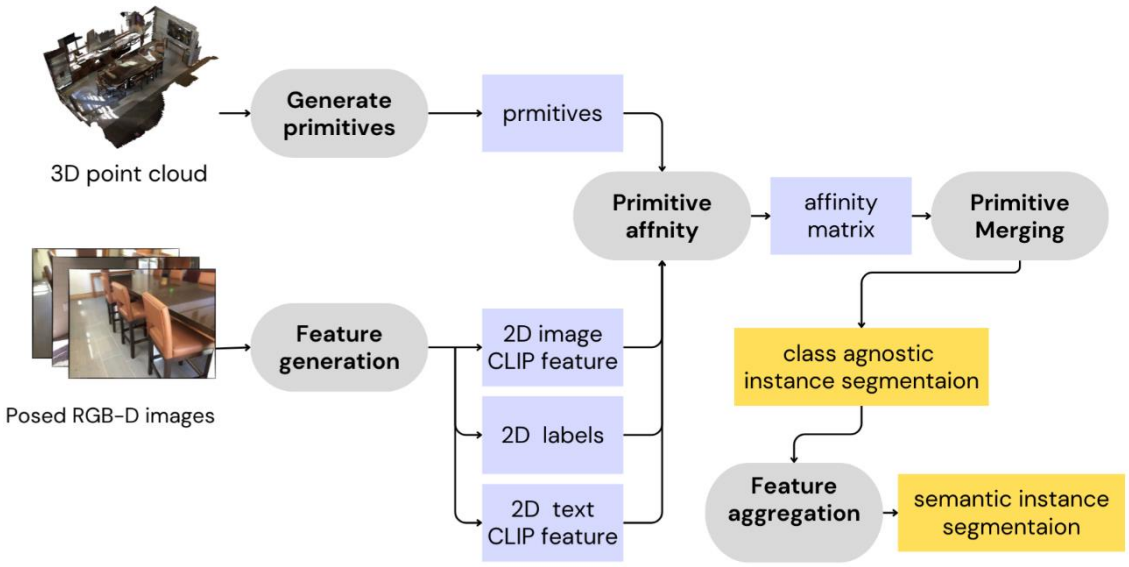

Figure 1. System Design. 粗體標示與下文標題對應。

2.2. Scene Graph Construction

Generate primitives. 跟隨 SAI3D[1]，使用 normal-based graph cut algorithm [3] 將 point cloud 切割成 primitives，以此將幾何性質相近的點集合在一起。

Features generation.

對於這三種特徵，此專題的核心想法是獲得每張圖片的 pixel-wise 特徵，而後利用 3D 到 2D 投影，就能得到每個 primitive 在各個影像上的特徵.以此作為相似度計算的根據。

(1) 2D indexed mask labels:

利用 Grounded SAM[8]中整合 RAM[10]的自動標註流程，實現每張 2D 圖片的 mask 分割，獲得每張圖片的 pixel-wise indexed mask label (例: 0、1、2)。

(2) 2D text CLIP feature.

Grounded SAM[8]中整合 RAM[10]的自動標註流程除了會給出2D mask 分割，也會給出每個 mask 的文字標籤(例: “chair”)，經由 CLIP text encoder 處理後得到2D mask-level 的 text CLIP feature. 配合 2D indexed mask labels ，即可得到

pixel-wise text CLIP feature。

(3) 2D image CLIP feature.

得到2D mask 分割後，我們可以進一步在原圖片上以每個 2D mask 的範圍作裁剪，並將裁剪後得到的圖片經由 CLIP image encoder 處理後得到 2D mask-level 的 image CLIP feature。 配合 2D indexed mask labels，即可得到 pixel-wise image CLIP feature。

另外本專題利用了 OpenMask3D[2]中的一個技巧，在裁剪時取3種不一樣的縮放比例，這三種比例的圖都會經過 CLIP image encoder，再將這個 feature 取平均，得到2D mask-level 的 image CLIP feature。在此專題中，該技巧大幅度強化了 image CLIP feature 的品質，使得 semantic evaluation 分數大幅上升。

2.3. Primitive Affinity

對於我們採用的三種特徵(indexed mask labels 、 image CLIP feature 、text CLIP feature)，我們都要計算對應的相似度分數。

Primitive projection.

對於第 p 個 3D primitive，透過對應的 camera 參數將其投影到第 m 個 2D 影像上，投影後的 primitives 在影像中可能部分可見，也可能完全被遮蔽。

我們定義其在該視角的投影可見性 $\text{seen}_{p,m}$為在影像 m 中可見的 3D 點比例，公式如下:

$$\text{seen}_{p,m} = \frac{1}{|\mathcal{I}_p|}\sum_{i\in\mathcal{I}_p} s_{i,m} \quad (1)$$

其中，$\mathcal{I}_p$為 primitive p所包含的 3D 點集合，$s_{i,m}$ 定義為:

$$s_{i,m} = \begin{cases} 1, \ if\ point\ i\ is\ visible\ in\ view\ m \\ 0\ , otherwise \end{cases} \quad (2)$$

若 $\text{seen}_{p,m}$ 為0，表示該 primitive 在此視角完全不可見，將其視為無效影像並捨棄。對於有效的影像，我們計算該 primitive 的投影特徵 $\text{feature}_{p,m}$ ，即該 primitive 在影像中所覆蓋的 2D 像素位置所對應的特徵平均：

$$\text{feature}_{p,m} = \frac{1}{|\mathcal{P}_{p,m}|}\sum_{u\in\mathcal{P}_{p,m}} \mathbf{f_{u,m}} \quad (3)$$

其中，$\mathcal{P}_{p,m}$為 primitive p 在影像 m 覆蓋的 2D 像素點集合，$\mathbf{f_{u,m}}$ 為影像 m 上位置u的 pixel-wise 特徵。

Affinity in a single view.

透過計算在每張影像兩個 primitives 的投影特徵之間的 cosine similarity 得到它們的 affinity score。

Affinity in multiple views.

由於 3D primitives 可以在不同的影像中被觀察到，它們之間的 affinity score 在不同視角下也會有所變化。將每個有效的 affinity score 視為候選分數，並使用投票機制來結合各個視角的結果，以達到跨視角的一致性。具體來說，對於兩 3D primitives Q_i 和Q_j，將其投影到 M 張圖片上(M 個視角)計算相似度分數$A_{i,j}^m$，SAI3D[1]為每個候選分數$A_{i,j}^m$計算一個權重，並使用加權和來得到最終的 affinity 分數$A_{i,j}$：

$$A_{i,j} = \frac{1}{\sum_{m=1}^{M}\omega_{i,j}^m}\sum_{m=1}^{M}\omega_{i,j}^m A_{i,j}^m \quad (4)$$

其中，第 m 個視角的權重$\omega_{i,j}^m$ 根據 Q_i 和Q_j 在第m個視角的投影: $Q_{i,m}$ 和$Q_{j,m}$ 的可見性計算，公式如下:

$$\omega_{i,j}^m = \frac{\sum_{p\in Q_i} 1(Valid(p,S_m))}{|Q_i|}\frac{\sum_{p\in Q_j} 1(Valid(p,S_m))}{|Q_j|} \quad (5)$$

其中，*Valid* 函數用來檢查投影點p 是否在 2D instance mask S_m 中可見。對於那些 primitives 不可見的影像，權重 $\omega_{i,j}^m$被設為 0，以排除無效的影像，確保分割的準確性。

2.4. Primitive Merging

基於 scene graph 和計算出的 affinity score matrix，SAI3D[1]使用 region growing 演算法來獲取最終的 3D class agnostic instance masks。這一算法會逐步合併具有高

affinity score 的 3D primitives。而我們沿用此一算法，並加入三種特徵的合併條件整合，使得三種特徵能相互輔助，生成品質更好的3D class agnostic instance masks。

Multi-level merging criteria：

在 region growing 過程中，每個 region 被表示為一個 primitives queue，將隊列頂部的 primitive pop 出，並用它來擴展相鄰的節點。在傳統的 region growing 算法中，如果 pop 出的 primitive 與鄰近節點具有較高的 affinity score，該節點就會被添加到區域中。然而，這種兩兩比較的方法容易產生錯誤，隨著增長過程的推進，這些錯誤會不斷累積。為了解決這一問題，SAI3D[1]提出了一種 Multi-level merging criteria，以層級方式計算 affinity score，將候選節點與區域內所有節點的 affinity 分數加總，並根據圖距離對其加權。

Progressive growing：

region growing 中的另一個重要參數是 affinity 閾值，它用來決定兩個區域是否應該合併。區域增長過程對閾值非常敏感，設置固定閾值通常會導致過分割或分割不足。基於這一觀察，SAI3D[1]提出了一種 Progressive growing 算法，將增長過程分解為多個階段，每個階段使用從高到低變化的閾值。在初始階段，使用嚴格的標準來合併小區域，防止錯誤的合併隨著增長過程積累。隨著區域逐漸增長為更穩定的較大區域，使用更寬鬆的標準來合併它們。這種動態閾值方法確保了合併框架能夠對連通性的逐步確定性保持敏感，從而提高了最終分割的準確性。SAI3D[1]原先只使用了 indexed mask labels 來計算 affinity score，而我們則用三種特徵計算了三種分數，並將合併的標準設計成當三個分數皆大於閾值才進行合併。Region growing 結束後即可得到 3D instance segmentation。

2.5. Feature aggregation.

最後我們為每個 3D instances 指派一個 CLIP 特徵以進行 semantic segmentation。利用**2.3. Primitive Affinity** 中建構的每個 primitive 在不同視角的 image CLIP 投影

特徵 feature$_{p,m}$ (我們使用三種特徵中的 image CLIP 特徵，以 img_feature$_{p,m}$ 表示) 以及可見性 seen$_{p,m}$。

首先，針對每個 primitive p，將其在所有視角中的投影特徵依照可見性進行加權平均，得到 primitive-level 的聚合特徵 agg_feature$_p$，公式如下:

$$\text{agg\_feature}_p = \frac{1}{\sum_{m=1}^{M} \text{seen}_{p,m} + \epsilon} \sum_{m=1}^{M} \text{seen}_{p,m} \cdot \text{img\_feature}_{p,m} \quad (6)$$

其中 ϵ 為防止除以零的微小常數，M 為影像總數。

接著，對於每個 instance（由多個 primitives 組成），我們進一步取其所包含的 primitives 的聚合特徵平均，得到 instance-level 的 CLIP 特徵表示：

$$\text{agg\_feature}_{\text{instance}} = \frac{1}{|\mathcal{P}_{\text{instance}}|} \sum_{p \in \mathcal{P}_{\text{instance}}} \text{agg\_feature}_p \quad (7)$$

其中$\mathcal{P}_{\text{instance}}$ 表示該 instance 所對應的 primitives 集合。

3. Experimental Results

3.1. Experiment Setup

Dataset. 我們的實驗使用了 ScanNetV2 與 ScanNet200 的 validation set。ScanNetV2 是一個針對室內場景的大型 3D 掃描資料集，涵蓋 20 個語義類別，廣泛用於 3D 語義與實例分割任務。ScanNet200 則是其擴充版本，包含更細緻的 200 個語義類別，挑戰性更高，更能檢驗方法在複雜語義空間中的泛化能力。

Evaluation metric. 我們使用廣泛採用的 Average Precision (AP) 作為數值結果的評估指標，我們回報以下情況的分數: 在 IoU 為 25% 和 50% 的情況下（AP@25、AP@50）以及在 50% 到 95% 區間內每隔 5% 的 IoU 所計算的平均值（即 mAP@[50:95]）。我們採用兩種評估設定方式：class-agnostic instance segmentation， 只關注 instance masks 本身的準確性; 以及 semantic instance segmentation，同時評估其對應的語義標籤。最終我們針對所有語義類別計算平均分數，以獲得整體的性能表現。

Baseline.我們將分數與 SAI3D[1]以及 SAI3D[1]的 baselines 做比較，來看我們的改進是否有增進它的效能。

Hyper parameters.我們使用五階段的 Multi-level region growing，三種特徵的合併閾值分別為: indexed mask labels, [0.9, 0.8, 0.7, 0.6, 0.5]、image CLIP feature, [0.98, 0.96, 0.95, 0.93, 0.92]、text CLIP feature, [0.98, 0.96, 0.94, 0.92, 0.91]。

3.2. Results

Fine-grained 3D segmentation.

Table 1.呈現了此專題在 ScanNetV2 資料集上進行 class-agnostic instance segmentation 的數值結果。我們的方法在所有評估指標中皆優於其他現有的方法，顯示出我們所做的改進的確產生了一個在open vocabulary class-agnostic instance segmentation 具競爭力的模型。

Table 1. Class-agnostic 3D instance segmentation on ScanNetV2 dataset.

粗體標示為該項指標的最高分數，我們方法的分數在三項皆達到最高。

Method	AP	AP_{50}	AP_{25}
HDBSCAN[11]	1.6	5.5	32.1
Nunes et al.[12]	2.3	7.3	30.5
Felzenszwalb et al.[13]	5.0	12.7	38.9
UnScene3D[14]	15.9	32.2	58.5
SAM3D[15]	20.2	34.0	53.3
SAI3D[1]	30.8	50.5	70.6
Ours	**33.8**	**54.2**	**71.1**

Open-vocabulary 3D object querying.

Table 2. 呈現了此專題在 ScanNet200 資料集上進行 semantic instance segmentation 的數值結果。雖然我們的 AP 指標略低於SAI3D[1]，但各項指標的表現已相當接近，顯示我們的方法在準確性上具有競爭力。我們的優勢在於對於此任務有更精簡的 end-to-end 流程，無須仰賴外接 Openmask3D[2] 等方法。我們也相信若進一步優化 hyperparameters，本方法有潛力能達到與SAI3D[1]持平甚至更佳的表現。

Table 2. Semantic instance segmentation on ScanNet200 dataset.

粗體標示為該項指標的最高分數，我們方法在三項均達到與 SAI3D 相近的分數。

Method	AP	AP_{50}	AP_{25}
OVIR-3D[16]	9.3	18.7	**25.0**
SAM3D[17]	9.8	15.2	20.7
SAI3D[1]	**12.7**	**18.8**	24.1
Ours	12.4	18.5	23.7

3.3. Ablation and Analysis

Merging logic for three features.

因為我們使用三種特徵算出三種分數，因此我們需要討論怎麼以這三種分數決定是否合併 primitives。我們首先只用 indexed mask labels 當作 merging 條件，來觀察他的 segmentation 品質。結果發現會有 under-segmentation 情形發生。以 **Figure 2**. 中為例，電視下的架子與垃圾桶被合併在一起，因此我們將合併條件設為需要同時滿足三種條件，利用另外兩個特徵來糾正 under-segmentation，在三種條件互相輔助下，**Figure 2.** 顯示成功分割出架子及垃圾桶。

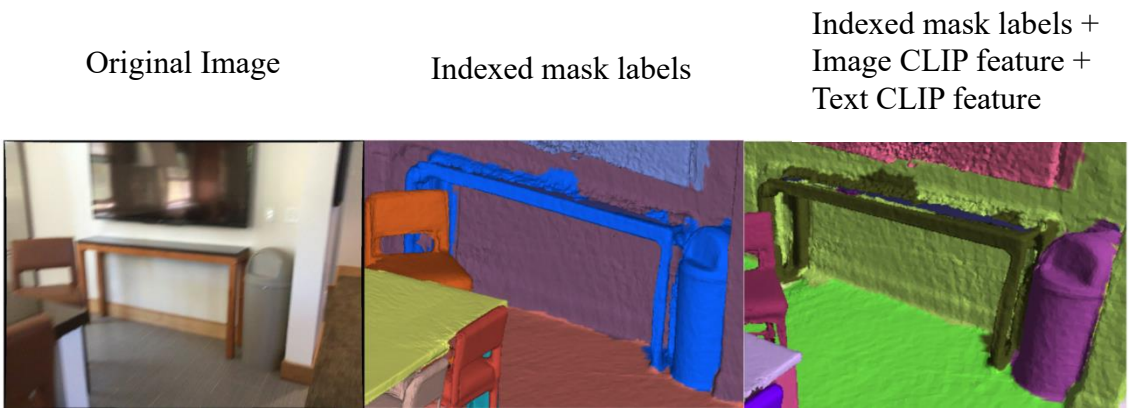

Figure 2. Under-segmentation example. 左邊為真實影像，中間為僅使用 indexed mask labels 特徵所生成的分割，右邊為使用三種特徵所生成的分割

4. Conclusion

我們提出了對 SAI3D[1]方法的改進，從一開始就將 CLIP feature 整合進流程裡，並利用其豐富的語義資訊協助分割任務，得到更好的3D class-agnostic instance

segmentation 能力。並且透過底層的特徵整合，使得流程更加的 end-to-end，無須外接 OpenMask3D[2]等方法即可完成 3D semantic instance segmentation 的任務，對 open vocabulary semantic instance segmentation 領域做出貢獻。

5. 心得感想

做專題的這一年多中，從一開始學習關於電腦視覺與機器學習的知識，並且讀相關的論文，再到後來開始提出專題 proposal，老師和學姐都幫助了我很多。給予我清晰的方向，讓我知道可以往哪個方向做探索以及嘗試。在確定方向開始實作後，更是每週跟我 Meeting，確定我現在的方向是對的，並且對實作細節給了我很多建議。非常感謝老師還有學姊們的幫助，才能讓我這個電腦視覺領域的初學者不斷進步，完成這個專題題目。在過程中，不斷地提出新想法以及做實驗去驗證的過程雖然很累，但做出預期中的結果時也很有成就感。專題不僅讓我增強了專業能力，也讓我累積了不斷試錯並修正的寶貴經驗，為將來的學習研究打下基礎。

6. Reference

[1] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. SAI3D: Segment any instance in 3d scenes. arXiv preprint arXiv:2312.11557, 2024

[2] Ayc¸a Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

[3] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. International journal of computer vision, 59:167–181, 2004.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and ´ Ross Girshick. Segment anything. arXiv:2304.02643, 2023.

[6] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In International Conference on Robotics and Automation (ICRA), 2023.

[7] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767, 2023.

[8] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, Lei Zhang. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159, 2024

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499, 2023

[10] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, Lei Zhang. Recognize Anything: A Strong Image Tagging Model. arXiv:2306.03514, 2023.

[11] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pages 33–42. IEEE, 2017.

[12] Lucas Nunes, Xieyuanli Chen, Rodrigo Marcuzzi, Aljosa Osep, Laura Leal-Taixe, Cyrill Stachniss, and Jens Behley. ´ Unsupervised class-agnostic instance segmentation of 3d lidar data for autonomous vehicles. IEEE Robotics and Automation Letters, 7(4):8713–8720, 2022

[13] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image

segmentation. International journal of computer vision, 59:167–181, 2004.

[14] David Rozenberszki, Or Litany, and Angela Dai. Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. arXiv preprint arXiv:2303.14541, 2023.

[15] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. 2023

[16] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In Conference on Robot Learning, pages 1610–1620. PMLR, 2023.

[17] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17619–17629, 2023.